DOCUMENT RESUME

ED 101 299 CS 001 590

TITLE	Design Concepts for a Measure of Effectiveness in Reading: A Feasibility Study.
INSTITUTION	Riverside Research Inst., New York, N.Y.
SPONS AGENCY	New York State Education Dept., Albany.
PUB DATE	Sep 73
NOTE	240p.
EDRS PRICE	MF-$0.76 HC-$12.05 PLUS POSTAGE
DESCRIPTORS	*Effective Teaching; Readability; *Reading; *Reading Achievement; *Reading Improvement; Reading Research; Reading Skills; Reading Tests; *Research Methodology; Research Needs

ABSTRACT

This feasibility study for developing a measure of effectiveness in reading contains five sections. "The Need and Requirements for a Measure of Effectiveness in Reading" presents the problem, functional specifications for a measure of effectiveness in reading, the minimum number of tasks required to uild an effectiveness measure, approaches to the measurement of effectiveness in reading, the Riverside Research Ins tute (RRI) approach toward the development of a measure of effectiveness in reading, and the RRI approach and the minimum work tasks for developing an effectiveness measure in reading. "Measuring Word Familiarity" discusses scales of word frequency, word frequencies and the lognormal distribution, construction of a word familiarity scale, and a familiarity-based vocabulary measure. "Measuring the Readability of English Texts" discusses the problem, a new readability formula, and construction of the RRI readability formula. "Implementing the Design Concepts in the Construction of Reading Tests" presents a plan for the construction of nonbiased tests and for computer-assisted tests. "Application of the Design Concepts for Qunatifying English Text in Setting and Monitoring Standards" discusses input data for setting standards and analysis of effective data. (WR)

RIVERSIDE RESEARCH INSTITUTE

80 West End Avenue / New York, New York 10023 / (212) 873-4000

September 1973

DESIGN CONCEPTS FOR A MEASURE OF EFFECTIVENESS IN READING

A FEASIBILITY STUDY

Prepared for

The New York State Education Department
Under Contract NYS C65911

F/281-5-10-1

Contributors To This Volume

This project was directed by Bertram L. Koslin. The research on the readability of text, on setting standards of reading competence, and on developing reading tests, was performed by Sandra Koslin. The research on word familiarity and the lognormal model was performed by Cornelius C. Langley. Ben Josephson, Jr. also contributed to this work.

The contributions of T. Anne Cleary and Walter H. MacGinitie, who reviewed and made valuable suggestions in connection with certain portions of this work, are gratefully acknowledged.

Table of Contents

Table of Contents (cont'd.)

Table of Contents (cont'd.)

Table of Contents (cont'd.)

Table of Contents (cont'd.)

Table of Contents (cont'd.)

Figures

Tables

Chapter I.

The Need and Requirements for a Measure of Effectiveness in Reading

A. The Problem

New York State spends at least a billion dollars a year on reading related instruction.[1] This sum is appropriated in the belief that expenditures of this magnitude are required to provide quality education in reading to all students in the State. Quality education is currently defined primarily in terms of input factors, e.g., high quality is associated with high per capita costs, elaborate facilities, large numbers of personnel, and so on. In New York State, as elsewhere, the outcomes of instruction, i.e., whether and how well students are, in fact, learning to read, are not emphasized in definitions

---

[1] New York State spends approximately six billion dollars annually for elementary and secondary education (SED, 1973, p.9). On the hypothesis that this six billion is divided equally over grades K-12, $461.5 million is spent per grade. In the following table we have assumed that the proportion of instructional time devoted to reading related activities varies according to grade. Documentation exists to support the assumption of 31% for grades 1-3 (OEO# B005114).

Grade	% Time for Reading Related Instruction	Costs in Millions
K	20	$ 92.3
1,2,3	31	$ 429.2
4,5,6	20	$ 276.9
7,8,9,10,11,12	15	$ 415.4
		$1,213.8

The accuracy of this estimate may be subject to some question. For example, it might be argued that less time is actually

instruction--is not emphasized in definitions of educational quality. This emphasis on inputs rather than outcomes has been summarized in the 1972 General Information Yearbook published by the National Assessment of Educational Progress:

> The only available measures of educational quality resulting from this investment [of billions of dollars] had been based upon inputs into the educational system such as teacher-student ratios, number of classrooms, and number of dollars spent per student. The tenuous assumption had been that the quality of educational outcomes--what students actually learn--was directly related to the quality of the inputs into the educational system. No significant direct assessment of educational outcomes had been made. (National Assessment of Educational Progress, 1972, p.1)

1. Current measures of outcomes in reading. The National Assessment of Educational Progress, in the quotation reproduced above, notes that "no significant [RRI's italics] direct assessment of educational outcomes has been made." The outcomes of reading instruction have not been entirely ignored; adult literacy is surveyed periodically, and schools frequently administer reading achievement tests. However, as the following sections of the report will show, these procedures do not directly assess reading outcomes because they do not yield

---

devoted to reading-related instruction in grades 7-12 than we have estimated, making the cost of reading instruction less than $1.2 billion. On the other hand, it might be argued that, because of the drop-out rate in the secondary schools, the $6 billion costs are not distributed evenly over the grades, but rather that proportionately more resources are allocated to the elementary grades than to the secondary grades. If

information that will permit a determination of what students have learned as a consequence of receiving reading instruction, i.e., how well students actually read.

a. Adult literacy. At the national level, adult literacy has been used as a measure of one outcome of schooling. The assumption is made that the existence of a literate adult population is evidence that the educational system is working effectively. For this purpose, literacy commonly has been defined simply in terms of years of schooling: a person is considered literate if he or she has completed a specified number of years of formal education. The number of years of schooling taken to define literacy varies among government agencies, but is currently in the range of five to eight years.

This definition of literacy does not constitute an adequate measure of the outcomes of instruction. Knowing how long a person has been in school does not necessarily provide any information about that person's reading ability. Studies indicate that reading ability, measured using standardized tests,

---

this is the case, the estimate of total reading instruction costs would need to be revised upwards (because of the greater proportion of time given to reading instruction in the elementary grades). It might also be argued that the application of Title I and New York State Urban Education funds to the teaching of reading raises the total costs of reading-related instruction. While such arguments (or others) would alter estimates of reading costs, we believe that the calculation shown above is a conservative estimate of the annual cost of reading instruction.

is frequently three to four grades below the number of years of schooling that the persons being tested have completed. One study conducted in the Woodlawn area of Chicago found that, although more than 90% of the persons sampled had completed at least the sixth grade, over 50% of them proved to be functional illiterates on the basis of achievement test results (Hilliard, 1963). Therefore, reliance on grade-completion criteria to define literacy provides little if any useful information concerning the real consequences of educational programs on the reading ability of students.

b. Standardized test scores. School districts and state education departments typically measure the outcomes of reading instruction by administering standardized, norm-referenced reading achievement tests to students. Agencies of the federal government also appear to be moving toward the use of performance on such tests to define literacy. However, scores on norm-referenced tests are inadequate measures of the outcomes of instruction because they do not provide information concerning either the attainment of standards of reading competence or the acquisition of particular reading skills.

Grade norms are widely misinterpreted. It is widely believed that these norms define standards of reading competence for each grade, i.e., that they define an objectively-determined level of performance that all children in that grade should be able to reach. It is not generally understood that

a grade "norm" is defined simply by the average test score that a sample of students of a given grade level did achieve during standardization of the test. Grade norms are established without regard for the particular levels of reading competence demonstrated by students; they depend only on the observed distribution of test scores earned by subjects in the standardization sample. Thus, norm-referenced scores cannot be used to determine whether students meet performance standards in reading.

Norm-referenced test scores are not directly interpretable with respect to what students have learned and, consequently, they provide no direct indication of how well a student will perform on any reading tasks that may be encountered in everyday life. If a twelfth-grade student obtains a reading score that places him at the twelfth-grade norm, no conclusion can be drawn concerning his ability to cope successfully with the reading tasks that he will meet in the adult world. All that can be inferred from this score is that his reading performance, compared with the performance of others in his age group, is about average on a particular set of test items. Perhaps the average twelfth-grade reader can read most of the adult materials that he will encounter, but there is no inherent property of the set of test items or of the test score that supports this conclusion. Being average, or even above

average, in relation to one's peers is no guarantee of competence on specific reading tasks.

It is not surprising, then, that the quality of education is currently defined primarily in input terms, considering that the common definition of literacy only takes account of how long a person has been in school rather than measuring his or her reading capability, and considering that norm-referenced tests only discriminate between different persons' performances on a non-generalizable set of tasks rather than providing a directly interpretable measure of reading skills. If the quality of education is to be defined in terms of outcomes, a new and different measure of reading ability is required.

2. The justification for developing a new measure of effectiveness in reading. The development of a new measure of effectiveness in reading can be justified in several ways. First of all, it can be justified in terms of the need for documented answers to several important questions that cannot be addressed substantively until a new measure of reading ability is available. An outcome measure of reading ability is needed to evaluate the different methodologies used to teach reading (e.g., different ways of organizing curricula and sequencing instructional activities) in terms of their long-term effectiveness. Furthermore, a new measure is needed to give concrete meaning to the phrase "equal educational

opportunity" through a study of the ultimate consequences of different programs in which resources are applied to overcome socioeconomic class differences among students.

The most important justification for developing a new effectiveness measure, however, is that it is urgently required to give substance to two important public processes in education: system accountability, and the allocation of resources (i.e., budget-making).

Effectiveness measures are an essential component of accountability processes in any field. During the last decade, much has been written and said about the need for system accountability in education. Public discussion of the matter has centered on two aspects of accountability. First, there has been a demand for demonstrated results from educational programs. The satisfaction of this demand requires measures that clearly show what students have learned as a consequence of receiving instruction. Second, the public has asked educational professionals to affirm with their constituents the specific educational objectives they have chosen to pursue and the means they are using to reach these objectives. The public is especially anxious to receive explanations for failures. To meet these demands, there is a need to document the relationship between alternative programs (clearly defined

with respect to objectives, methodology, implementation, and so on) and their measured effectiveness.

The need for effectiveness measures is equally critical in budget-making. The public budget-making process in education results in a series of resource allocation decisions. There is never enough money available in education to do all the things that educators or the public wish to do. Therefore, decisions must be made to spend the money (i.e., to allocate the available resources) on one set of educational programs rather than another. If such decisions are to be made rationally, they must be based on the expected, measurable effectiveness of different educational programs relative to their costs.

Since the effectiveness of an educational program can only be judged in terms of what it has actually accomplished (that is, in terms of what students have learned), public resource allocation processes cannot take place rationally unless and until there are effectiveness measures available that provide directly interpretable data demonstrating what students have learned from different instructional programs. Furthermore, since budget-making is a public process, it would be desirable to present program effectiveness information in a form that citizens can readily understand, thus facilitating their informed participation.

B. The Functional Specifications for a Measure of Effectiveness in Reading

The properties that are desired in a measure of effectiveness in reading constitute a set of functional specifications. These functional specifications are as follows:

1. The capability to measure individual reading effectiveness. Since education is concerned with the development of individuals, the measure must yield reliable individual scores of reading comprehension.

2. The capability to measure system effectiveness. It must be possible to aggregate the scores of individuals (by grade, sex, ethnicity, etc.) to determine how well the educational system is performing for different target groups in different schools, districts, regions, and statewide.

3. The capability to measure progress toward adult reading competence. The test must be able to measure the progress of individuals (and groups) toward becoming competent adult readers.

- It must measure the ability to cope with societal reading requirements imposed by law, such as comprehending income tax forms or drivers' license applications, and with other materials intended by government agencies for the protection and well-being of citizens.

- It must measure the ability to read materials necessary to enter various vocations or professions.
- It must measure the ability to read materials that enable individuals to function competently in their own behalf, such as advertisements, insurance policies, repair manuals, etc.

4. The capability to measure growth in reading ability. The measure should be able to detect small changes in reading ability, such as might be expected to occur in one year's time. Measurement of group growth is an essential requirement of the measure. Measurement of individual growth, if feasible, is highly desirable.

5. The capability to measure reading ability over the entire school age range. Continuity of measurement, beginning in the primary grades, is necessary for measuring progress toward adult competence and for detecting growth. Therefore, the measure should be applicable over all or nearly all of the public school age range.

6. The capability to furnish meaningful scores. Scores on the measure should be readily and accurately understood by persons without technical knowledge of statistics or test construction procedures, such as parents, legislators, teachers, etc. Therefore, it must be possible to present scores in terms that are meaningful to such persons without sacrificing precision in reporting.

C. The Minimum Number of Tasks Required to Build an Effectiveness Measure

There is a minimum number of tasks that must be executed in order to construct a reading effectiveness measure.

1. Legislative-political tasks. The task of building an effectiveness measure logically requires a clear statement of the objectives that the educational system is trying to achieve. Therefore, it is desirable that the persons who are empowered to do so define the standards or expectations of reading competence.

The actual setting of standards (a matter of value judgment) is outside the province of science; rather, it is the job of government. However, scientists can contribute sound, impartial technical work to describe adult reading requirements, and to define and analyze the consequences of alternative standards, so that government can choose among alternatives as rationally as possible. Since reading demands (and language) change over time, and since students entering school need to be prepared to meet the reading requirements that they will face as adults approximately 15 years later, the analytic work carried out by scientists should include some amount of forecasting.

2. Scientific-technical tasks. Whether or not formal standards are established, the scientific and technical tasks to be carried out in building a measure of effectiveness in reading remain essentially the same. The technical task, however, is simplified somewhat when standards have been set and measurement need only determine whether or not those standards are met. These tasks are as follows:

- To define adult reading tasks. Identify the various kinds of materials that adults are called upon to read.
- To scale adult reading tasks. It is reasonable to assume that the number of adult reading tasks will be too large to test students' ability on all of them. The large number of adult reading tasks suggests that an approach which treats reading tasks individually will be less productive than one which scales reading tasks according to the extent to which they share one or more properties. Reading tasks with similar scale values can be clustered into groups. With the tasks organized or clustered in groups, performance on a given task would allow valid inferences to be made about an individual's performance on any other task within the same group.

- To define "reading comprehension." At the outset of test construction, it is necessary to define the construct "reading comprehension," i.e., to specify the cognitive skills to be encompassed by this construct, so that appropriate test items may be chosen. Furthermore, criteria of comprehension must be specified. These criteria define the test performance to be accepted as evidence that a student satisfactorily comprehends what he has read.
- To carry out the technical development of the test:
  - To select item formats;
  - To demonstrate construct validity (that is, having defined the construct "reading comprehension," to demonstrate that the tests used are valid measures of this construct); and
  - To determine test reliability (and to develop new procedures for calculating reliability, if needed).

D. Existing Approaches to the Measurement of Effectiveness in Reading

In recent years, there have been a number of large-scale reading projects related to the measurement of effectiveness in reading. In the following sections, several of the more important efforts will be examined and reviewed in relation to the functional specifications and minimum work tasks outlined above.

1. The National Assessment of Educational Progress. Reading is one of ten subject areas covered in the National Assessment of Educational Progress (NAEP) currently being conducted under the auspices of the Education Commission of the States. The purpose of NAEP is to collect census-like data on a nationwide basis concerning the educational achievement of Americans in selected content areas. The NAEP's plan calls for periodic retesting to detect changes in achievement.

When the decision was made to undertake NAEP in reading, panels of reading specialists, educators, and test developers were convened to define reading objectives that would represent "a set of goals which are agreed upon as desirable directions in the education of children" (National Assessment of Educational Progress, 1970, p.2). The draft objectives agreed to by the panelists were submitted to groups of lay citizens to ensure that the objectives to be measured would be perceived as

important by the public. After the reading objectives were decided upon, professional item writers prepared test "exercises" to measure those objectives in four age groups: 9, 13, 17, and 26-35.[2] Each objective was measured in all age groups, but the test items differed by age, since a decision had been made to try to keep the median percentage of success at 50% per objective per grade. During 1970-71, 500 test exercises were administered to approximately 100,000 subjects in the four age groups.

1.1 NAEP's measure and the functional specifications for an effectiveness measure in reading. Although NAEP is an ambitious undertaking that provides a great deal of descriptive data about the reading achievement of students and young adults, it does not meet all the functional specifications for an effectiveness measure in reading.

a. Individual scores. NAEP does not provide individual scores.

b. System scores. NAEP does not provide scores for schools, districts, or states, though such data could

---

[2] The objectives are: to comprehend what is read; to analyze what is read; to use what is read; to reason logically from what is read; and to make judgments about what is read. Another objective--to have attitudes about and an interest in reading--was agreed upon, but was not assessed at all in the first national assessment of reading.

presumably be provided if needed. NAEP does provide data for various regions of the country and by various types and sizes of communities.

c. Measurement of progress toward adult competence. NAEP provides no means for measuring progress toward adult competence. Since adult reading competence levels were not defined, progress toward such competence logically cannot be measured. (The reading objectives that are measured pertain to desirable skills that any reader should have; no distinctions are made between objectives for various age groups.)

d. Measurement of growth. NAEP cannot measure growth in reading for any individual or group, since there is no known relationship between the difficulties of the exercises used in the tests constructed for the different age levels. Some exercises were administered at two or three age levels, but only a small number were administered at all age levels. Since the exercises were not scaled for difficulty, differences in scores on different tests administered over time to the same students are uninterpretable with respect to growth.

e. Applicability over age range. While NAEP covers a wide age span, the use of different exercises that have no known relation to each other in tests for different age groups raises doubts as to whether the measurement of achievement can be considered continuous over the age range.

f. Interpretability of scores. It is NAEP policy that results be reported in a way that will be understandable to educators and interested citizens. Therefore, the statistical presentation is kept quite simple. However, because the data are primarily reported by individual "exercise" per age level (and for demographic subgroups within each age level) the reader must synthesize a great deal of detailed information. The interpretive load on readers remains large even when the "exercises" are grouped by "themes" (sets of "exercises" clustered for reporting purposes) and by objectives.

1.2 NAEP and the minimum work tasks required to develop an effectiveness measure. NAEP has carried out only some of the minimum work tasks required to construct an effectiveness measure.

a. Input for setting standards. NAEP provides no input data for policy-makers to set standards.

b. Define adult reading tasks. NAEP does not define adult reading tasks; instead, reading objectives are defined as cognitive skills that any reader should have.

c. Organize or cluster reading tasks. NAEP does not cluster reading tasks in the process of constructing tests. However, following the administration of reading assessment measures in 1970-71, the test items themselves were organized for reporting purposes into clusters that "have something in common."

d. Define the construct "reading comprehension." NAEP does define the cognitive skills to be measured in a test

of reading comprehension. However, criteria of comprehension are not specified.

e. Construct validity. The construct validity of the measures has not been demonstrated, but their content validity has been established through an elaborate review procedure.

f. Reliability. No information on reliability has been provided yet.

2. The ETS adult reading tasks. One goal of the Targeted Research and Development Project in Reading, sponsored by the U.S. Office of Education, is construction for ten-year-olds of a criterion-referenced test that will predict competent performance on adult reading tasks "selected to have favorable returns to the individual and to society in general" (Educational Testing Service, 1971, p.2). Educational Testing Service (ETS) is assembling the set of adult reading tasks that will serve as criterion for the test.

To define representative adult reading tasks, ETS conducted a survey of what they have termed a "national probability sample" of adults to learn about their daily reading habits. In this survey, respondents described the types of reading done during a 24-hour period, the amount of time devoted to each type, and the importance of each type. Prototype reading tasks were built to represent the main types of reading activities reported in the survey and regarded as important by the respondents. The survey data and prototype

tasks served as input to expert panels that were to study and rank these tasks on a scale of benefits to the individual and society, and to suggest high-benefit reading activities that were inadequately represented in the set.

The ETS plan calls for administering the tasks to a national sample of adults in order to determine task intercorrelations and thus to find, through factor analysis, dimensions of reading competence. A second sample, on which demographic data will be collected, will be used to establish relationships between performance on various tasks and economic, social, and cultural status levels. The tasks that are finally chosen on the basis of the field tests will serve as the criterion that the proposed test for ten-year-olds will eventually have to predict.

Although the ETS work is not yet complete, it may be tentatively reviewed in relation to the functional specifications and minimum work tasks required for an effectiveness measure.

2.1 ETS' and OE's proposed measures and the functional requirements for an effectiveness measure in reading.

a. Individual and system scores. Both types of scores presumably could be obtained.

b. Measurement of progress toward adult competence. Yes, but only in a limited sense, namely, whether students at

ten years of age show satisfactory performance (yet to be defined) on test items that predict criterion performance as adults. The nature and extent of the relationship between test tasks and criterion tasks are not yet specified.

c. Measurement of growth. There is no provision for the measurement of growth in reading competence.

d. Applicability over age range. The test is intended only for ten-year-olds.

e. Interpretability of scores. Unknown at this time.

2.2 The status of the ETS work and the minimum tasks required to develop an effectiveness measure.

a. Input for setting standards. Establishing links between success on reading tasks and the educational or economic status of adults should provide useful input for those empowered to set standards.

b. Define adult reading tasks. This has been done by professional test developers and panels of expert advisors, taking into account the results of a national survey of reading habits.

c. Organize or cluster reading tasks. Not clear. ETS does plan to factor analyze performance on tasks to identify underlying dimensions of reading competence. This analysis may do more for defining the construct "reading comprehension" than for organizing the tasks themselves.

d. Define the construct "reading comprehension." The proposed factor analyses, c. above, should help to define the skills to be measured in a test of comprehension. Criteria of comprehension have not yet been specified.

e. Construct validity. Construct validation of the criterion adult tasks is planned on a national sample. Plans for determining the construct validity of the actual test for ten-year-olds have not yet been reported.

f. Reliability. Not yet determined.

3. The Harris surveys of "survival" requirements in reading. Louis Harris and Associates have been commissioned by the National Reading Center to conduct periodic surveys to determine how well adults are able to carry out reading tasks of the type required to "survive" in contemporary American society. The surveys focus on practical reading skills required to cope with common experiences in the lives of Americans, such as following directions for direct dialing of telephone calls, understanding employment and housing advertisements, responding appropriately to questions on application forms, and so on. Test items directly measuring the ability to carry out tasks such as these are administered in individual interviews to a national sample of respondents selected to represent the civilian non-institutional population of the United States. Results are reported in the form of a composite index of

reading difficulty, calculated by weighting items according to their difficulty. This index is to be used on a regular basis as a measure of functional reading problems in the United States (Harris and Associates, 1971).

While the Harris surveys provide useful information concerning selected functional reading skills of adults in various demographic groups, they do not meet several important functional specifications for a measure of reading effectiveness.

3.1 The Harris "survival" measures and the functional requirements for an effectiveness measure in reading.

a. Individual scores. Individual scores can be provided.

b. System scores. There is no readily identifiable "system," other than the nation's schools as a whole.

c. Measurement of progress toward adult competence. The Harris surveys are not designed to measure progress toward adult competence.

d. Measurement of growth. The Harris surveys do not measure growth in reading achievement.

e. Applicability across age range. The Harris surveys are designed only for persons 16 years old or older.

f. Interpretability of scores. Reports of the percent of respondents answering various numbers of test items

correctly should be understandable to persons without technical training. However, the National Difficulty Index is not easily understood.

3.2 The Harris surveys and the minimum tasks required to develop an effectiveness measure. The Harris surveys have carried out some but not all of the minimum tasks required to develop a measure of reading effectiveness.

a. Input for setting standards. It is uncertain whether or not the Harris approach provides useful input data for setting standards of reading competence.

b. Define adult reading tasks. Harris has used expert opinion to define a restricted set of reading tasks, namely those considered essential for "survival."

c. Organize or cluster reading tasks. Harris has not organized or clustered the reading tasks in any way.

d. Define the construct "reading comprehension." Uncertain.

e. Construct validity and reliability. No information provided.

4. The Adult Performance Level Study. The purpose of the Adult Performance Level Study (APLS), being conducted at the University of Texas with the support of the United States Office of Education, is to define literacy operationally in terms of reading and other skills required to function

effectively in "areas of need" which are important for survival in our society. Six areas--occupational knowledge, consumer economics, health, community resources, government and law, and transportation--were identified by means of literature reviews, surveys of professional opinion, conferences on adult needs with lay and professional participants, and interviews with undereducated persons. In each area, reading (and other[3]) skills required for effective functioning were listed. Criterion-referenced test items built to test these reading behaviors were validated in a nationwide study by determining the relationship between success on test items and several indicators of the economic and educational status of respondents. Based on analyses of field test data, a revised list of adult performance requirements was developed.

The APLS plans to increase the comprehensiveness of its coverage and conduct more extensive validation studies. The APLS expects the set of functional reading (and other) tasks that will be compiled to serve to guide the content of courses in adult basic education, and also to serve as a means of assessing functional literacy.

---

[3] Other skills measured are writing, speaking or listening, computation, problem solving, and interpersonal relations.

Although the APLS research is still in process, published reports concerning progress and research plans (Adult Performance Level Project Staff, 1973) have enabled RRI to review tentatively the extent to which APLS is likely to yield reading tests that meet the functional specifications for an effectiveness measure.

4.1 The APLS measures and the functional requirements for a measure of effectiveness in reading.

a. Individual scores. No information provided. However, with data reported on an item-by-item basis, there is no obvious basis for a meaningful summary score.

b. Measurement of progress toward adult competence. APLS does not provide for the measurement of progress toward adult competence.

c. Measurement of growth. APLS does not measure growth in reading achievement.

d. Applicability over age range. APLS is designed for adults only.

e. Interpretability of scores. Uncertain at this time. However, the plan to report data on an item-by-item basis poses problems of summarizing data.

4.2 APLS and the minimum tasks required to develop an effectiveness measure in reading. The APLS has carried out some, but not all, of the minimum work tasks required to construct an effectiveness measure.

a. Input for setting standards. The analysis of important functional reading skills and the relating of performance on reading tasks to economic and educational status should constitute useful input to government for setting standards.

b. Define adult reading tasks. This has been done through a combination of reviews of research, expert opinion, and surveys of adult (lay) opinion.

c. Organize or cluster reading tasks. APLS groups reading tasks by "areas of need" and "objectives." The tasks themselves are treated individually and have not been scaled. Although no empirical evidence is given to support the claim, APLS contends that performance on particular tasks is predictive of performance in the entire "area."

d. Define the construct "reading comprehension." The reading skills to be measured have been defined. However, criteria of comprehension have not been specified.

e. Construct validity. APLS is establishing the construct validity of tasks by determining whether predicted relations are obtained between performance on reading tasks and the economic and educational status of respondents.

f. Reliability. No information has been provided yet.

5. Summary of the approaches taken to measure effectiveness in reading. In Section D, four major approaches to the measurement of reading competence were reviewed, both with respect to how well the functional specifications for a measure of effectiveness in reading were met, and with respect to whether or not the minimum work tasks required to develop such a measure were undertaken. The review showed that each approach meets some of the functional specifications and that, in each case, some of the necessary work tasks have been completed, but that none has met or completed all of them.

None of the approaches reviewed is capable of meeting two of the functional specifications, namely those for the measurement of growth in reading achievement and the measurement of progress toward adult reading competence, although meeting these specifications is critical to the measurement of both individual and system reading effectiveness. With respect to the minimum work tasks, none of the approaches has yet successfully solved the problem of scaling adult reading tasks. Failure to organize adult reading tasks creates obvious difficulties in building a test that adequately samples the task domain, and in reporting and interpreting data.

RRI recognizes that, since none of the projects reviewed set out originally to develop measures of individual and system reading effectiveness suitable for use over a wide

age range, they should not be faulted for failing to do so. The point of the review has been to show that, as innovative and useful as these national efforts at assessment and measurement are, they do not completely satisfy the requirements of public education for a reading effectiveness measure.

E.' The RRI Approach Toward the Development of a Measure of Effectiveness in Reading

Under the terms of a contract with the New York State Education Department,[4] RRI formulated design concepts that would contribute in a significant way to setting standards of reading competence and that would lead to the construction of a reading effectiveness measure meeting the functional specifications outlined earlier. Under the same contract, RRI also developed a plan for implementing these design concepts, i.e., for performing the scientific and technical work required to develop a new measure of reading effectiveness.

The heart of the RRI approach, and what distinguishes it from other attempts to measure reading competence, is an emphasis on finding ways to characterize adult reading materials quantitatively, and to use these quantitative properties of reading materials both as inputs for setting standards and as the basis for determining whether those standards are being met (i.e., for test design). RRI reasoned that, if reading materials can be quantitatively scaled in terms of significant variables, standards can be defined in terms of the scale values found for selected adult reading materials, and that

---

[4] Contract #C65911.

competence can be assessed by determining a person's ability to read materials having those specified scale values. RRI further reasoned that, if ways could be found to describe all reading materials quantitatively, it would become possible to relate performance on any reading task to performance on other reading tasks and thereby to open the way to measure growth in reading competence.

Three major design concepts for quantitatively characterizing reading materials were explored. Following a detailed evaluation, two of these, word familiarity and the readability of text, were found to be powerful enough to establish the feasibility of a single effectiveness measure that will meet all of the functional specifications described earlier in this chapter. These two design concepts and their applications are developed in detail in this report.

The third concept, syntactic complexity of text, was judged to be potentially valuable but of doubtful practical utility in the near term. At present, models of the syntactic structure of the English language do not appear to be sufficiently developed to permit reliable scaling of the complexity of large samples of English text.[5] Therefore, syntactic complexity was dropped as a design concept.

---

[5] Reviews of the pertinent literature led RRI to conclude that, at the present time, syntactic models of English do not appear

1. Word familiarity and readability. Studies (cited in Chapter IV) of reading test performance suggest that reading success depends on two principal factors: knowledge of individual word meanings; and comprehension of connected text. RRI therefore reasoned that, if the vocabulary and the textual characteristics of reading materials could be scaled, it should be possible to define adult reading competence logically in terms of the vocabulary that a reader must know and the difficulty of text that he must comprehend to be able to read adult level materials competently.

---

to be sufficiently developed to permit reliable mechanical scaling of the complexity of passages. Although sentence structures can be reliably parsed by applying transformational and other contemporary theories of grammar, much uncertainty remains concerning sentence complexity.

At least part of the problem stems from the fact that the way in which syntactic and semantic factors interact to produce complexity for the reader is not yet understood. The scaling of passage complexity is also restricted by the fact that syntactic analysis has largely been limited to single sentences. Consequently, the study of factors producing complexity across sentences in connected prose has barely begun.

RRI's conclusions on these matters are supported by a paper on the scaling of syntactic and semantic complexity prepared for SED by Finn (1973) in an effort independent of the work described in this report. After reviewing his own and others' work in some detail, Finn concludes that the application of syntactic models to written passages is years away.

Lacking formal models of passage complexity, RRI considered the possibility of analyzing certain syntactic features of individual sentences, and of averaging over sentences to derive summary syntactic complexity scores for passages as Chomsky (1971) has done. Unfortunately, any such analysis would have to be carried out by hand by a trained grammarian, since an

RRI proposes that the vocabulary of reading materials be characterized according to the familiarity of different words to readers of the language. In this work, word familiarity is defined in terms of the different frequencies of occurrence of words in written English (frequently occurring words are taken to be more familiar than infrequently occurring words). Passages of text differ from one another in the proportion of words of high, low, and moderate frequency that they contain. Some passages contain many common or very familiar words; others contain a large proportion of rare words.

---

adequate job cannot be done by available computer parsing programs. Chomsky (personal communication) has described this hand-analytic procedure as " . . . cumbersome and time consuming, and probably not worth all the effort that it requires." In view of the very large quantity of material that will probably need to be scaled in constructing the RRI reading effectiveness measure (see Chapters IV and V), hand analysis of syntax must be ruled out for practical reasons.

If an intensive programming effort were undertaken, RRI might be able to achieve computer analysis of syntax. However, such an intensive effort does not appear to be justified in terms of the additional knowledge that would be gained about students' reading ability. As the review in Chapter III will show, syntactic factors are so entwined with readability that separately evaluating students' ability to comprehend materials at different levels of syntactic complexity would probably be redundant with evaluating their ability to comprehend materials at different levels of readability.

If the word-frequency characteristics of adult reading materials can be quantitatively described, it should be possible to define reading competence operationally by estimating the number of words from different frequency bands that a person would need to know to be able to read those materials successfully.

RRI further proposes that connected text be scaled for its comprehension difficulty or readability. Readability is a summary characteristic of text determined by the interaction of structural and stylistic factors. These factors combine to make some passages of text easier (or harder) to comprehend than others. By scaling the readability of adult reading materials, the level of textual difficulty that a person must be able to comprehend to be able to read specified adult materials successfully can be determined.

Neither the scaling of readability nor the measurement of word frequencies are new ideas. To the best of RRI's knowledge, however, they have never before been used separately or together as a means either for defining standards of reading competence or for measuring the extent to which those standards have been met.

2. Defining standards of competence. The information obtained from systematic, quantitative measurements of word familiarity and readability can be used to set standards of adult reading competence. The same capability that enables

the familiarity of words occurring in any passage of text and tne difficulty of that passage to be measured quantitatively also makes it possible to define quantitatively the skill levels--in terms of word knowledge and comprehension--required to perform any reading task. If government specifies the written materials that graduates of the educational system are expected to be able to read, and if these written materials are scaled for readability and word familiarity, then the levels of word knowledge and comprehension skill that graduates of the educational system must reach are operationally defined by the measured word frequency and readability characteristics of the designated materials. Thus, performance standards for educational processes can be established.

3. A new measure of reading competence. A new approach to the construction of an effectiveness measure in reading follows logically from the preceding argument that standards of reading competence can be defined in terms of the scaled, linguistic properties of reading materials. Once quantitative standards of reading competence are defined empirically, simple and direct measurements of the extent to which students have attained these standards can be made by administering reading tests consisting of items that have been scaled for the same linguistic properties that were used to define the standards, i.e., for word familiarity and readability.

In the comprehension sections of such a test, some passages would correspond in readability to the difficulty level designated as the adult standard. Other passages would be more and less difficult. A student's performance on passages whose readability corresponds to the standard of adult difficulty is directly interpretable in terms of whether or not adult standards of competence have been met. In the event that adult standards have not been specified when the tests are administered, performance on passages of text scaled with respect to readability can be used to provide an accurate assessment of the level of difficulty of text that a graduate is able to comprehend.[6]

The vocabulary sections of the test of reading effectiveness would contain words sampled systematically from the different word frequency bands constructed from RRI's analysis of adult reading materials. The performance of students on these sections of the test could be interpreted directly in terms of their knowledge of words in each of the frequency bands, and student performance could be evaluated, in this way, in terms of the vocabulary required to meet adult reading

---

[6] Such information cannot be obtained from current norm-referenced tests because the readability of passages in those tests is not systematically varied.

standards. In the event that standards have not been specified when the tests are administered, a student's knowledge of words can be compared to the vocabulary used in a wide range of adult materials.[7]

If the vocabulary and passages of text that make up the reading effectiveness test reflect the full range of school reading materials--from primers all the way to adult-level text--then the test instrument could be administered periodically to monitor a student's progress toward adult reading competence as the student moves through school.

The word familiarity and readability design concepts lead not only to measurement of the progress toward and attainment of adult reading competence, but also to measurement of students' attainment of grade-level reading objectives. Following the same logic that was used earlier to define adult competence, grade-level objectives in reading can be defined operationally by analyzing instructional materials to determine the expectations that are being placed on students concerning knowledge of words of particular frequencies of occurrence and comprehension of text of given difficulty at each

---

[7] Such information cannot be obtained from current norm-referenced vocabulary tests because the so-called "blueprints" for such tests do not involve the systematic selection of test words from different frequency bands.

grade level. With grade level objectives for word familiarity and reading comprehension thus defined, it would be possible to determine the extent to which students meet grade-level expectations with the proposed measure of effectiveness in reading, i.e., the extent to which students at different grade levels know words of appropriate frequencies of occurrence and can comprehend text of suitable levels of readability.

F. The RRI Approach and the Functional Specifications for a Measure of Effectiveness in Reading

RRI's approach to the definition and measurement of effectiveness in reading, outlined above, will result in a measure of reading competence that will meet all the functional specifications listed earlier in this chapter.

1. Individual scores. The RRI effectiveness measure will yield *word knowledge and reading comprehension scores* for each individual. Depending on whether or not standards of competence have been established, scores may be used either to evaluate achievement levels in relation to grade-level objectives or adult standards of competence, or to describe the current achievement level of the student.

2. System scores. The scores of individuals on the RRI effectiveness measure can be aggregated to obtain reading effectiveness scores for schools, districts, regions, or for the state as a whole. System effectiveness could be examined for various subgroups (e.g., by sex, ethnicity) by aggregating over appropriate individuals.

3. Measurement of progress toward adult competence. The effectiveness measure can be used to determine the levels of skill and knowledge required to perform important adult reading tasks competently, e.g., to read materials required by law, to read materials required for entry into different

vocations, and to read materials required to function as a competent adult. Since readability and word familiarity can be scaled continuously, progress toward becoming a competent adult reader in any or all of these areas can be monitored by periodically administering tests in which vocabulary and readability gradually reach the level appropriate to adult tasks.

4. Measurement of growth. The scaling of reading materials makes it possible directly to compare students' performance on different forms of the same test or on a graduated series of tests, thereby providing a basis for measuring growth in reading achievement. Reliable measurement of group growth should be obtainable using scaled materials. Reliable measurement of individual growth depends essentially on whether sufficient time can be devoted to testing in narrow word-frequency and readability ranges (see Chapter IV).

5. Applicability over age range. Use of the word frequency and readability design concepts as a basis for tests of competence makes it possible, in principle, to use a common measurement scale over the entire public school age range. However, in the earliest grades, different reading programs may not overlap sufficient y in vocabulary to provide a common base for testing. Thus, the earliest grade in which a common measure can be used that is unbiased with respect to any particular reading program must be determined.

6. Interpretability of scores. The RRI effectiveness measure will yield scores that should be readily understood. Reports of reading competence will be anchored directly to performance, rather than to how a student's reading achievement compares with that of his peers.[8]

A report on reading comprehension might read, "John can comprehend all the materials used in his fifth grade reading program. A sample paragraph is enclosed illustrating the most difficult materials used in this program. His present reading skill would probably allow him to comprehend adult reading materials such as (examples given)." Vocabulary test results can also be reported in a way that can readily be understood by parents, teachers, and other concerned citizens. A report might read, "John meets all of the vocabulary requirements of his fifth grade program. Relative to what he will need to know as an adult, John now knows 80% of common words, 40% of moderately familiar words, and 15% of rare words."

---

[8] Peer comparisons, e.g., percentile and stanine scores, can be provided if needed.

G. The RRI Approach and the Minimum Work Tasks for Developing an Effectiveness Measure in Reading

In addition to developing the design concepts introduced above and a general strategy for their implementation, RRI has also begun, or has developed programmatic plans for, the minimum work tasks required to produce a reading effectiveness measure.

1. Legislative-political tasks. The RRI design concepts will make it possible to provide government with precise information concerning the readability and word familiarity characteristics of various adult reading materials. This information will make it possible to define reading competence operationally in terms of the levels of readability that a student must comprehend and the word knowledge he must have in various familiarity bands to read competently those materials designated by government as essential or important.

2. Scientific-technical tasks.

- Define adult reading tasks. Alternative ways of defining a representative collection of adult reading materials were examined. A preliminary decision has been reached to use the domain of periodicals to define the range of content and difficulty of adult reading materials. Several smaller domains of practical importance have also been identified.

- Cluster adult reading tasks. Representative random samples of text from periodicals will be scaled for word familiarity and readability. Any specific reading task or set of tasks can be similarly scaled and related to the text drawn from the periodicals.
- Define the construct "reading comprehension." In building the reading effectiveness measure, RRI will define reading comprehension in terms of a student's ability to understand a passage of text sufficiently well to correctly identify words that have been deleted from it. Thus, RRI has decided to measure one general comprehension factor rather than several distinct comprehension subskills. Since research indicates that comprehension subskills are highly interrelated, measurement of a general comprehension factor should adequately measure the skill(s) usually encompassed by the construct "reading comprehension." To be credited with comprehending a passage, a student must correctly answer a sufficient number of questions to reduce the probability that his score occurred by chance along to an acceptably low level.
- Technical development of measures. Although actual developement of the measures has not yet begun,

some plans for their technical development have been formulated.

- Item selection. A quasi-cloze format has been chosen for items in the comprehension section of the test. In this format, a word is deleted from text and students must select the deleted word from among several options provided. Response options will be controlled for word familiarity and semantic plausibility. In addition, a strategy has been formulated for regulating the familiarity of response options in the vocabulary section of the test, so that test results can be interpreted unambiguously with respect to a student's knowledge of words in various frequency bands.
- Demonstration of construct validity. A preliminary strategy has been formulated for verifying that the proposed items adequately measure reading comprehension. In essence, validation would be carried out using factor analytic techniques to compare test results obtained using the RRI items with results obtained when other item types are used and when multiple-comprehension subskills are measured. The validity of the criterion established for crediting a student with comprehension of material at a given

level of readability can be tested in separate studies requiring behavioral evidence that the student can comprehend other materials of the same difficulty level.

- Determination of test reliability. RRI believes that it will be necessary to employ or develop new procedures for determining test reliability. Procedures currently used for determining the reliability of norm-referenced tests will not be applicable to RRI's criterion-referenced effectiveness measure.

H. Other Applications of the Design Concepts

The principal applications of the readability and word familiarity design concepts are to provide information for the setting of standards of reading competence and to measure the extent to which students have met those standards. Other important applications have also been identified. For example, RRI believes that the design concepts could be used to analyze instructional materials to determine whether the vocabulary and readability demands that are placed on students at different grade levels contribute to the failure to meet adult standards of reading competence. This line of investigation could lead to recommendations for changes in the readability and vocabulary content of instructional materials. Since such recommended changes would be designed to make the readability and vocabulary content of instructional materials more rational, their implementation would increase the likelihood that students will reach adult standards of reading competence as they progress through the educational system.

Chapter II

Measuring Word Familiarity

It is well known that the words in the English lexicon, like those of other languages, can be classified in terms of their frequencies of occurrence, i.e., some words occur more often than others. It seems reasonable to assume that a competent adult reader must know all of the most frequently occurring words, nearly all of the next-to-most frequent words, somewhat fewer of the words in the next lower frequency class, etc.

Therefore, it follows that, if the frequency-of-occurrence characteristics of words in adult materials can be scaled, and if students' knowledge of words in various frequency bands can be systematically determined, it should be possible to compare a students' word knowledge with the vocabulary required to read adult materials competently.

The idea of formally scaling words in terms of their frequencies of occurrence in order to measure students' vocabularies represents a significant departure from current practice in testing, but one that is needed if a new effectiveness measure is to be built. Tests of word knowledge are, of course, widely used both in norm-referenced measures of reading achievement and in measures of general ability (I.Q.). However, no direct inference concerning the scope or size of a student's

vocabulary can be drawn from such tests of word knowledge because the words tested in norm-referenced measures are not obtained from a systematic sampling of the words in the lexicon. Systematic sampling is not required in norm-referenced tests because the purpose of such tests is primarily to determine how many words (of those tested) the child knows compared with other students. With this purpose in mind, words in the final version of a norm-referenced test are apt to be chosen largely for their power to discriminate among students.

Therefore, currently used measures of word knowledge do not provide a basis for judging whether a student's knowledge of words is adequate to allow him to read adult materials. However, for the measurement of effectiveness in reading, a test is required that permits direct inferences concerning students' progress toward and attainment of adult standards of reading competence. RRI believes that the concept of word frequency or familiarity[1] can lead to the construction of a measure from

---

[1] In this report, the terms word frequency and word familiarity are used interchangeably. It is intuitively reasonable to suppose that words which occur very frequently in written language will be very familiar to readers, while seldom used words will be less familiar. Research supports the assumption of such a relationship between frequency and familiarity. It has been found, for example, that words with high frequencies of occurrence are recognized more rapidly (Howes and Solomon, 1951), and heard more readily in noise (Postman and Rosenzweig, 1957) than words with low frequencies of occurrence.

It has also been found that reading rates are faster for more frequent words than they are for less frequent words (Pierce

which such inferences are possible. The concept of word familiarity permits the vocabulary characteristics of adult materials to be scaled quantitatively, and provides a basis for building tests to measure students' word knowledge on the same scales.

---

and Karlin, 1956). In short, experimental subjects behave as though they are, in fact, more familiar with words having a high frequency of occurrence than they are with words that occur less often.

A. Scales of Word Frequency

1. The need to scale word frequencies formally. The frequency of words must be formally scaled. While the notion of word familiarity is easily understood in a rough, intuitive way, it is quite another matter to apply it with any precision. Everyone, for example, will agree that the word arachnid is much less familiar than the word house; but it is not so clear how the comparison would go between more closely-matched pairs of words like philanthropy and extrapolation, or table and book. To make objective rankings of the familiarity of closely-matched words, a numerical scale of word familiarity is required.

Such scales can be built by drawing a large sample of written material and observing the number of times particular words occur in it. The result of dividing the number of occurrences of a particular word by the total number of words in the sample gives the observed frequency of occurrence of that word, which may also be taken to define its familiarity. For example, suppose that in a sample of 1,000,000 words, the word water occurs 1,500 times. Then the observed frequency (and the familiarity) of the word water[2] in that sample would be calculated as:

---

[2] An extremely common word has been chosen for this example; most words have very much lower frequencies. To avoid the nuisance of working with very small numbers, it will probably be desirable to alter the definition somewhat, say, by using a logarithmic scale or some such device.

$$\frac{1,500}{1,000,000} = 0.0015$$

2. Existing scales of word frequency. A number of word frequency scales, based on word counts of samples of the English language, have been developed. The best known and most extensive of these is the Teacher's word book of 30,000 words (Thorndike and Lorge, 1944), which is based on a count of over 20 million words from a variety of printed sources. A more recent count has been made by Kucera and Francis (1967), but is based only on one million words.

Counts based on specialized materials have also been prepared. These include Horn's (1926) count of five million words in personal and business correspondence, Rinsland's (1945) count of six million words in children's compositions, Howes' (1966) count of 250,000 words of adult spoken English and, most recently, the American Heritage Intermediate Corpus (Carroll, Richman, and Davies, 1971) of five million words in instructional materials used in grades three to nine. All the existing scales, however, are based on samples that are too small to allow the precision of measurement required to draw accurate inferences concerning word knowledge in various frequency bands, and to detect growth.

3. The need for an enormous word sample. The vast majority of English words have extremely low frequencies. To obtain a reasonably accurate estimate of these frequencies, and enormous

sample is required. The need for large samples can be illustrated by the following data from the American Heritage Intermediate (AHI) corpus of five million words.

The following words (among others) each occurred exactly five times in the AHI corpus: helm, cuticle, gossamer, boredom, villa, grate, cutlass, stuffy, repast, debut, jocund, gadfly, therapeutic, sabotage, euglena, decoy. The observed frequency for each of these words is 0.000001 (once per million). The following words (among others) each occurred exactly ten times in the AHI corpus: esophagus, mermaid, gadget, lavish, needy, plaintiff, lilac, vengeance, sustain, musty, belfry, rascal. The observed frequency for each of these words is 0.000002 (twice per million).

However, it cannot be stated with certainty that words in the second set are more familiar than those in the first set, even though they exhibit twice the frequency in the AHI corpus, because it is not certain that the observed frequency of any of these words (obtained from this particular sample) equals the true frequency of the words in the entire universe of written English. The small size of the sample results in uncertainty in true frequency estimations. For example, if a word occurs five times in a sample of five million words, there is a 95% certainty that its true frequency of occurrence lies somewhere between 0.00000042 and 0.00000238 (that is, 95% of the words which occur five times in the sample have true frequencies

between these limits). For a word which occurs ten times in the sample, the corresponding limits are 0.00000107 and 0.00000373. The relatively small size of the sample produces "fuzziness," or low precision, in the estimate, and overlap between the 95% confidence limits for the two sets of words.

Now observe the effect of increased sample size on the true frequency estimates of the words considered above. In a sample of fifty million, the 95% confidence limits for a word that occurs fifty times would be 0.00000075 and 0.00000133; those for a word occurring one hundred times would be 0.00000163 and 0.00000244. The overlap is gone, but there is still some fuzziness. In a sample of five hundred million words, we get much better resolution: the 95% confidence limits for a word occurring five hundred times are 0.00000091 and 0.00000109, while those for a word occurring a thousand times are 0.00000138 and 0.00000213.

No matter how large the sample is (within practical limits), there will always remain a rather large class of words for which only a crude estimate of true frequency will be possible. These words are the so-called hapax legomena, the very rare words which occur only once or twice even in an extremely large sample. For such words it is impossible to obtain good resolution of their true frequency of occurrence, as shown in Table 1.

Table 1

95% Confidence Limits for True Frequencies of Words Occurring 1, 2, 3, and 4 Times in a Sample of Size N (where N is assumed to be very large).

No. of Occurrences	95% Confidence Limits for True Frequency	
1	$\frac{0.172}{N+4}$	$\frac{5.828}{N+4}$
2	$\frac{0.536}{N+4}$	$\frac{7.464}{N+4}$
3	$\frac{1.000}{N+4}$	$\frac{9.000}{N+4}$
4	$\frac{1.528}{N+4}$	$\frac{10.472}{N+4}$

There will also be a sizeable class of words which do not occur at all, even in an extremely large sample. Although these words cannot possibly be identified, the data of Carroll et al. (1971) suggest that there may be a way to estimate their total number. In the case of the AHI corpus of five million words, for example, it was estimated that only about 15% of all English word types were represented.

B. Word Frequencies and the Lognormal Distribution

While RRI cannot make direct use of existing word counts in establishing a word familiarity scale because they are based on samples that are too small, such counts do provide valuable information concerning the general shape and properties of word frequency distributions.

1. Types and tokens. In describing word frequency distributions, it will be helpful to introduce two terms commonly used in vocabulary studies. A type is a particular word, while a token is a particular occurrence of a word. For example, the AHI corpus consists of five million tokens, representing some eighty-five thousand different types. The single type water accounts for about 7500 tokens (i.e., this word occurred 7500 times) in the AHI corpus.

2. The shape of word frequency distributions. Now let us suppose that we have a sample of N tokens, representing G different types. If a particular type occurs J times in the sample, then the fraction J/N is the observed frequency of that type. Of course, there may be many types having the same observed frequency. Let G_J denote the number of types which have an observed frequency J/N. Then the fraction $\left(G_J\right)/G$ is the proportion of types having this observed frequency. If we plot the fraction $\left(G_J\right)/G$ as a function of observed frequency, we will get a curve similar to that in Fig. 1.

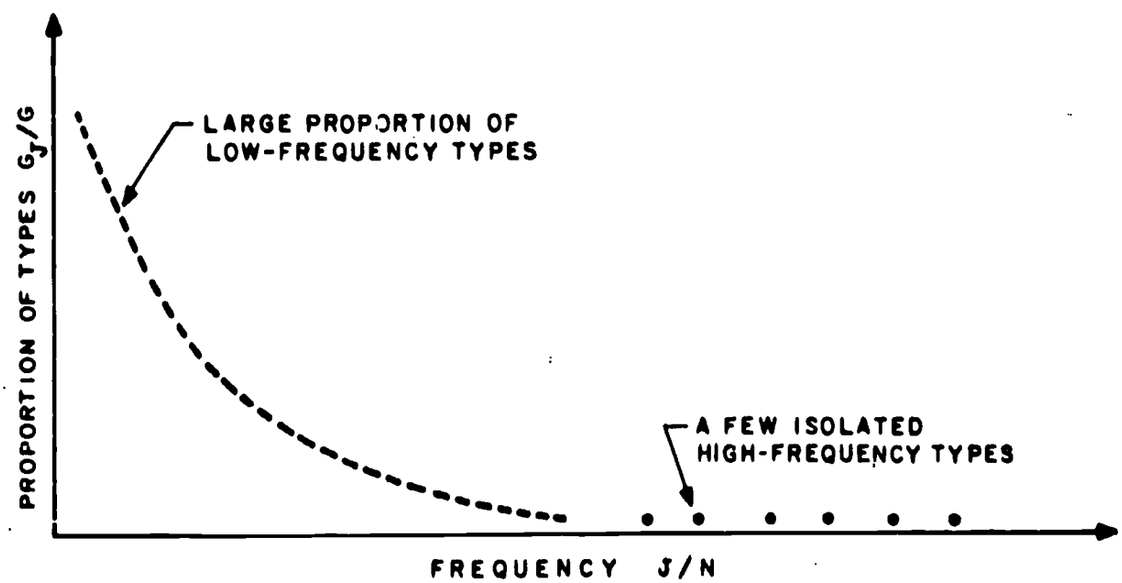


FIG. 1 TYPE DISTRIBUTION CURVE FOR A MODEST-SIZED SAMPLE.

The curve shown in Fig. 1 is typical of word frequency data obtained from samples of modest size. It shows that, the lower the frequency, the more numerous the types. If the sample is extremely large, however, the picture changes: The curve bends downward at the left end as a consequence of the fact that the number of very rare types is outnumbered by the number of types of slightly higher frequency. Fig. 2 shows how the curve would look for a very large sample.

The reader should note that the curve in Fig. 2 has been distorted in order to show its shape more clearly. If the curve were drawn accurately to scale, the peak would be extremely close to the vertical axis, and the curve would be very narrow and sharply bent on the high-frequency side. Fig. 3 shows the shape of the curve more accurately, but even it is distorted to some extent: the peak is not as close to the vertical axis, and the curve is not as narrow and sharply bent, as they would be if the curve were drawn to scale.

From such diagrams, we may begin to develop a sound intuition about the make up of the English lexicon. It consists of a small number of extremely common types, plus an enormous number of types having very low frequencies. In the AHI sample, for example, the ten most common types (the, of, and, a, to, in, is, you, that, it) accounted for nearly 25% of all the tokens, while the hundred most common types accounted for nearly 50% of them. At the other extreme, there were more than 35,000 types

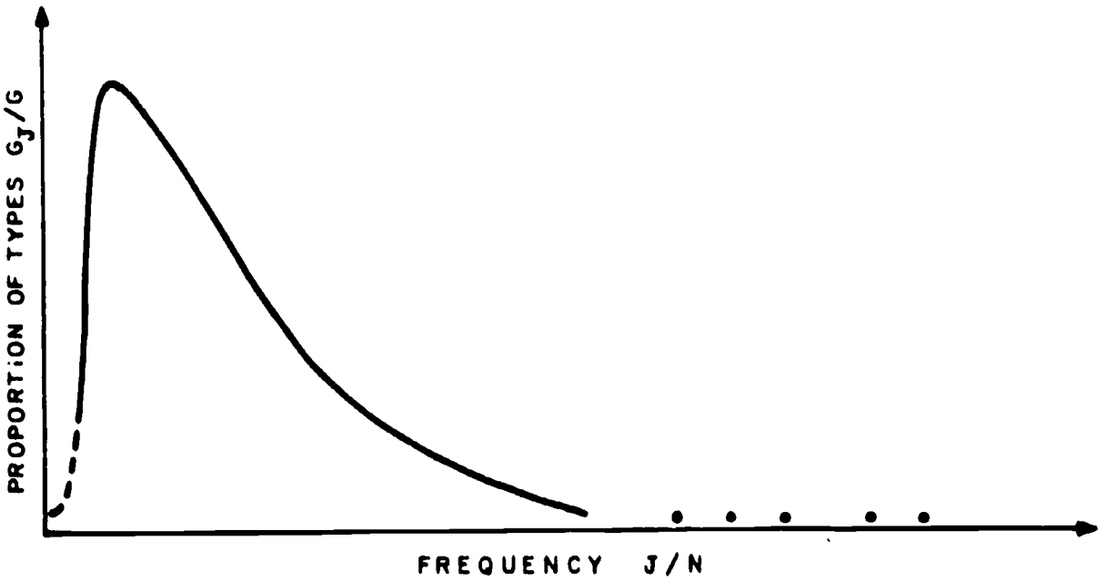


FIG. 2 TYPE DISTRIBUTION CURVE FOR A VERY LARGE SAMPLE, DRAWN TO DISTORTED SCALES TO SHOW GENERAL SHAPE.

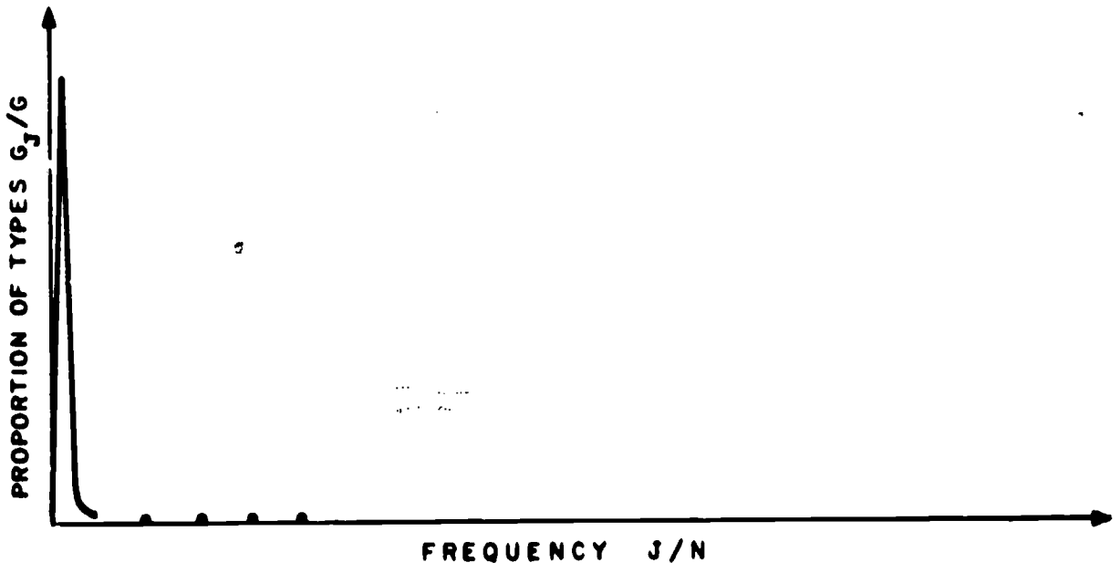


FIG. 3 TYPE DISTRIBUTION CURVE FOR A VERY LARGE SAMPLE, DRAWN MORE NEARLY TO SCALE THAN FIGURE 2.

which occurred only once each, and another 12,000 which occurred only twice each. These figures probably would be somewhat different for a lemmatized corpus,[3] but the general pattern should be much the same.

3. The lognormal distribution. Several investigators (Herdan, 1960; Carroll, 1968) have found that these type distribution curves are matched with extraordinary precision and faithfulness over most of their range by what are known as lognormal distributions. This name is derived from the fact that, if a logarithmic scale rather than a linear scale is used for the horizontal "frequency" axis, then the type distribution curve assumes the familiar, symmetrical shape of the normal distribution.

The fact that the lexicon is lognormally distributed is extremely important because the lognormal model furnishes a way to describe type distributions precisely and succinctly. Any normal distribution is completely described by two parameters, μ and σ (the mean and the standard deviation). The same two parameters are sufficient to describe the corresponding lognormal distribution (although they no longer have quite the same meaning). In other words, the lexicon can be characterized with

---

[3] Lemmatization, discussed later in this chapter, refers to the process of reducing a word to its dictionary form by stripping it of affixes. For example, teach, teacher, and teaching would all be reduced to teach and thus would be treated as one type for purposes of counting frequencies of occurrence. No lemmatization was carried out in constructing the AHI corpus.

precision if μ and σ can be calculated. This calculation involves some fairly sophisticated mathematics, the details of which are of no concern here. The interested reader will find an outline of this calculation in Appendix A.

The significance of the lognormal model for the development of a reading effectiveness measure is that the model permits written materials to be described quantitatively in a compact form. The ability to quantify and to represent simply the familiarity characteristics of word samples makes it possible to compare easily the vocabularies in different kinds of written materials. Such comparisons are important in setting standards of adult reading competence (see Chapter V). The ability to describe vocabulary quantitatively in a compact form also makes it feasible, if samples of words are properly drawn and students' knowledge of these words is appropriately tested, to measure the extent to which a student has attained the vocabulary required to read adult materials.

C. The Construction of a Word Familiarity Scale

The construction of a word frequency scale is straightforward, at least in principle. It begins with the specification and acquisition of a representative collection, or corpus, of English reading materials. The domain of periodicals appears to be representative of the universe of written English, and has been selected tentatively as the domain from which a representative, random sample will be drawn to constitute the RRI corpus. The RRI word familiarity scale will be developed from the words that occur in this corpus (see Chapter IV).

The corpus will be constructed by drawing a random sample of passages of text from different periodicals. The amount of text drawn from any periodical will be proportional to its circulation or press run. The representative, random sample of passages of text will be fed into a computer, and the occurrence of various word types will be counted to yield a scale of word familiarity. Once the corpus of periodicals has been scaled and μ and σ have been calculated, smaller specialized corpora (e.g., instructional materials, government publications) can also be scaled, and the results related to the results obtained for the major corpus.

1. Some problems to be resolved in counting word types. In order to count the occurrence of word types, a set of decisions must be made concerning how certain word types will be treated.

a. Inflections. Perhaps the most important decision is that associated with grammatical inflections. For example, should the words talk, talks, talking, talker, and talked be regarded as distinct words having different familiarities, or should they all be grouped together under the single word talk? This process of classifying words into "dictionary entries" is known as lemmatization.

Although final decisions on lemmatization have not yet been made, it is likely that only regular grammatical forms will be lemmatized. RRI is inclined to believe that irregular forms must be learned as separate vocabulary items, and do not inherit the same familiarity status as their roots. Once the decisions concerning lemmatization strategy have been made, it is expected that the great bulk of the lemmatization can be carried out economically by computer. There will be some need for human editing, however, to catch such mistakes as, say, lemmatizing stocking together with stock, or hammer together with ham.

b. Compound words. Closely akin to the lemmatization problem is the problem associated with compound words. The word coffeepot, for example, occurs far less often in print than either of its two components. Therefore, it would receive an unduly low familiarity rating unless some special attention is paid to it. It may well prove reasonable to assign to such

compounds the same familiarity rating as their least familiar components. Such decisions will be postponed until the data have been gathered.

c. Affixes. The situation is even more puzzling in the case of words affected by prefixes or suffixes, for these addenda themselves are not equally familiar. One may be willing to agree that a person who understands the word legal will also understand the word illegal, but the same agreement would not be forthcoming in the cases of words like quasi-legal or extra-legal. Another difficulty arises from the fact that it is entirely possible for the affix-laden word itself to be more familiar than the root word from which it was (presumably) derived, e.g., uncanny, unkempt, unravel.

d. Spelling. Still another nuisance is the incidence in English of variant spellings. It would seem reasonable to treat minor variants (like color and colour) as if they were identical words, but it is not so clear what should be done with cases like jail and gaol. Similar remarks would also apply to the deliberately misspelled words that occur in renderings of dialect speech (or in the poetry of Ogden Nash).

e. Shortened forms. Contractions and abbreviations also present problems. In some cases, they should probably be classified as separate words. However, what is to be done with the longer and rarer ones, such as shouldn't and Phila? Ought

they to be treated as separate words, or do they belong with should and Philadelphia insofar as familiarity is concerned?

f. Homographs. Thus far, consideration has been given to the need to put certain non-identical words into the same pigeonhole for the purpose of establishing their familiarities. The opposite problem must also be confronted: separating identically-spelled words having different meanings.

The simplest example of this problem is furnished by homographs, i.e., pairs of words which, though totally different, are spelled alike: does (is doing) and does (female deer); or entrance (doorway) and entrance (bewitch). Unless some very special and elaborate precautions are taken, the computer, in its innocence, will certainly throw both members of such a pair into the same bin, thus arriving at a misleading count.

Sometimes capitalization provides a sufficient clue for separation. Thus, we can distinguish Polish (the language) from polish (shine) and March (the month) from march (step). But in other cases there seems to be no way to effect the separation except by examining the context of the word each time it occurs in the sample.

g. Multiple meanings. A more subtle, and much more common, source of false collocation of words is the fact that one and the same word may be used with two or more different

meanings. Consider, for example, the word rather in the following two phases:

(1) a rather tall building

(2) I'd rather stay home.

It would be gratuitous to assume that rather in (1) is just as familiar as rather in (2). For all practical purposes, the problem here is the same as that of outright homographs.

From the foregoing discussion, it might appear that, in order to arrive at a sound word-familiarity scale, not only must RRI draw an immense sample of words, but also the context in which each of these words occurs must be scrutinized. However, the task is not as forbidding as it may seem.

First, only a small minority of words exhibit truly different multiple meanings. These words can be identified in advance by consulting a dictionary. Only these words would call for examinations of context. Second, it would not be necessary to go through the entire sample to determine the frequency of each sense of an ambiguous word. A few dozen citations should be sufficient to establish a statistically stable pattern. To illustrate, let us take the word entrance. Suppose that, of the first 100 occurrences of this word in the sample, it is found, by examining context, that 85 have the sense "doorway" while 15 have the sense "bewitch." Then it is probably safe to assume that 85% of all occurrences of this word have the first sense, and 15% the second.

Finally, it is possible that the entire task can be done by computer. The recent work of P.J. Stone and his colleagues at Harvard has resulted in a sophisticated computer program which reputedly can make decisions concerning which sense of a given word is intended in a passage of text. This program is available and if the cost of using it is not too great, it can be used to dispose of virtually all complications that arise, including the fact that words of identical appearance can differ in meaning depending on context.

h. Technical terms. One further complication arises in connection with rare words. Many of these words are highly specialized technical terms, and there is a tendency to use such words repeatedly, if at all. An example is the word lathykin, which is the name of a special tool used in making stained-glass windows. If this word appears at all in the sample, it is likely to appear not just once but dozens of times because there is no other word for this tool. Thus we will tend to overestimate its true frequency in the universe of written English.

There are several ways to deal with this complication. One way would be simply to delete such words from the corpus. This procedure, however, would require human editing and would introduce an undesirable element of subjectivity. Another way would be to let them stand, relying on the rather low readership enjoyed by such words (i.e., the low circulation

of journals that use the word) to correct the overestimate of their frequencies. Still another way was followed in the analysis of the AHI corpus. It consisted of introducing, for each word, a measure of dispersion of usage of that word among the different subject-matter categories. If a word was found to be used only in a few subject categories, then the observed frequency of that word was reduced (in some cases considerably). RRI will probably use a similar procedure, although the exact form of the dispersion measure has not been determined yet.

D. A Familiarity-Based Vocabulary Measure

The construction of a measure of students' word knowledge in various frequency bands should be relatively straightforward, once a word-familiarit scale has been established. Since a profile of a student's word knowledge in several frequency bands is to be provided, test words should not be chosen at random from many different parts of the familiarity scale. Rather, they should be chosen carefully from a few specified, narrow intervals of the scale.

From a student's performance on the test words in a particular frequency interval, it should be possible to draw a valid inference concerning the student's knowledge of all the types belonging to that frequency interval. For example, suppose a subject gets 15 right out of 20 words tested in interval A. Then it can be stated with 95% confidence that he knows between 53% and 89% of the types belonging to A. Or, it can be stated with 95% confidence that he knows at least 57% of those types. If more than 20 items are tested in interval A, the precision of the estimates can be improved. For example, if the number of test words from interval A is raised to 100, and if the subject gets 75 of these right, then it can be stated with 95% confidence that he knows between 65% and 83% of the words belonging to interval A, or that he knows at least 67% of these words. Assuming that the number of test items remains the same,

such precise conclusions would not be possible if the selection of test words had not been restricted to a narrow interval on the word familiarity scale.

By testing a student's knowledge of words in each of several narrow frequency bands, several different scores can be obtained. These separate scores would form a profile that would describe the student's vocabulary, relative to what is required to read adult materials competently, with far more precision than any single score could hope to do.

Chapter III

Measuring the Readability of English Text

A. Problem and Background

It is a commonplace observation that some passages of written material are harder to understand than others. This chapter deals with design concepts for scaling this variability in passage difficulty, or readability. As the term is generally used, readability refers to the relative comprehensibility of different textual materials: if Passage A is easier to understand than Passage B, A is said to be more readable than B. Scaling of the readability of written materials is an essential prerequisite for the development of the proposed RRI reading effectiveness measure.

1. Need for a measure of readability. A measure of readability is prerequisite to the measurement of reading effectiveness. It provides a means of objectively quantifying standards of reading competence. Once readability has been scaled, levels of difficulty that characterize the materials a competent adult must read can be specified. By testing students' ability to read materials at those levels of difficulty, it is possible to determine whether or not they have met adult standards. Moreover, by administering

tests calibrated for increasing passage difficulty (from the simplest levels up to adult levels), students' progress toward adult reading competence can be measured.

Scaling the difficulty of reading materials is also a prerequisite to analyzing how well the short-term performance objectives of a reading program have been met. The capability to quantify the reading difficulty level of instructional materials used in various grades and programs makes it possible to test students on passages of difficulty comparable to that found in the instructional materials being used, and thus permits a determination to be made of whether or not the expected level of reading competence has been achieved.

The designers and publishers of standardized norm-referenced reading tests provide no information concerning the readability of any of the passages of text used in the tests. Since reading comprehension levels on norm-referenced tests are defined exclusively by comparing a student's performance with that of his peers, the precise readability of passages included in the tests does not really matter. The only significant property that these passages and questions must have is that they span a sufficiently wide range of difficulty to permit reliable ranking of children's relative reading comprehension skill. Average performance on a set of test passages by children of a given age defines the norm of reading achievement for that age, regardless of the particular

characteristics of the passages or questions on which the scores were obtained. For this reason, as noted in Chapter I, a reading achievement score on a norm-referenced test cannot be interpreted directly in terms of the kinds of material a child can read and comprehend. By contrast, since the specified purpose for building the RRI reading effectiveness measure is to be able to interpret reading test scores directly in terms of the kinds of materials students are able to read, the readability characteristics of the test passages used for measuring achievement must be known precisely.

2. A brief history of readability research and formula construction. The idea of scaling the difficulty level of reading materials is not new. Klare (1963) may be stretching the point when he traces interest in the comprehensibility of messages back to Biblical days and to the Talmudists of the Middle Ages. However, it is certainly true that since the 1920's there has been a steady stream of studies concerned with rating the relative comprehensibility of different reading materials and with identifying the various structural and stylistic factors that make one passage relatively more or less difficult to understand than another.[1]

---

[1] The discussion of readability in this chapter is limited to analyses of structural and stylistic variables that account for differences in the comprehensibility of text. Subsidiary aspects of readability such as type face and legibility, or the extent to which the readability of a passage varies across readers as a function of interest or experience, will not be discussed. This limitation is imposed because the objective of RRI's work is to scale the language characteristics of the reading materials themselves.

The primary purpose of readability research in education has been to find a simple and objective way to judge the appropriateness of different reading materials for students with a given level of reading ability, without going through the actual process of asking students to read the materials to determine which are too easy and which are too difficult. A secondary purpose, more typical of publishing and journalism than of education, has been to learn how to write or revise materials so that they meet appropriate levels of difficulty for specified audiences of readers.

The desire for a quick way to judge the appropriateness of materials for particular groups of students has led to the construction of "formulas" that typically estimate the approximate grade-level reading achievement needed to comprehend the materials. Many readability formulas have been proposed during the last 50 years. The total number published is not known, since it depends on how "formula" is defined. In his review, Klare (1963) lists 31 formulas, though more than 50 can probably be cited if less stringent rules are applied. Some of the better known and more widely used formulas were developed by Lorge (1939), Dale-Chall (1948), Flesch (1948), Gunning (1952), Farr-Jenkins-Patterson (1951), and Spache (1953). Simplified formulas to facilitate more rapid calculation of readability have recently been published by Fry (1968) and McLaughlin (1969).

Most readability formulas have been built in the following way. The author of the formula begins by selecting some set of passages, known to vary in difficulty, to serve as his criterion scale of readability. He counts the occurrence in those passages of structural or stylistic variables that he believes cause some passages of text to be more difficult to comprehend than others. He then calculates the algebraic combination of those variables that best predicts the (predetermined) difficulty of the criterion scale. The equation giving the best prediction becomes the formula.

Although most formulas are alike in that they were built by weighting predictor variables against a criterion scale of reading difficulty, they vary widely in almost every other significant way. They differ with respect to the factors used to predict readability, the criterion scale against which the formula was originally validated, the difficulty range to which the formula is applicable, the definition of comprehension used in scoring the original criterion passages, the sampling method for selecting passages of text for analyses, the counting rules used in computation, and the units in which readability is expressed.

3. Shortcomings of existing formulas. Readability formulas have been widely used by publishers to control or adjust the difficulty of instructional materials. They have also been extensively used by educators, who employ them to

decide whether instructional materials are suitable for students who have a given level of reading ability. However, existing formulas may not be accurate enough to warrant the wide use they receive, and they certainly do not appear adequate to the task of building the RRI measure of reading effectiveness.

One serious shortcoming of existing formulas is that they do not predict enough of the variance in the criterion scale scores that they were originally designed to predict. The Dale-Chall formula, which is reported (Klare, 1963; Powers, Sumner & Kearl, 1958) to have the highest validity coefficient (r = .71) of any wide-range formula, is able to account for only about one half of the variability in reading difficulty of the criterion passages. Other popular formulas are even less powerful predictors of criterion readability.[2]

A second shortcoming of available formulas is that they are not very accurate. The Dale-Chall formula, which is reported to have the smallest standard error of measurement

---

[2] An exception is the Spache formula that, with a validity coefficient of r = .82 (Spache, 1953), accounts for about two-thirds of criterion readability variance. However, the Spache formula is suitable only for primary-grade reading materials, and was built using a criterion scale (publishers' grade level assignment of texts) which leads, for technical reasons, to inflated estimates of validity.

of any wide-range formula (Powers, Sumner & Kearl, 1958), has an error of measurement of .77 grades. This means that over 30% of the time Dale-Chall readability scores will deviate from "true" readability scores by more than three quarters of a school year. Other formulas either have larger errors of measurement or report none at all.

Where only rough estimates of readability are required, perhaps this relatively low level of precision in the formulas can be tolerated, although some critics contend that current readability formulas do more harm than good because of their imprecision (Bormuth, 1966). However, the proposed reading effectiveness measure requires greater precision in the scaling of passage difficulty than the available formulas provide, in order to be able reliably to detect small gains in reading achievement (such as might occur from the beginning to the end of a school year) and to measure accurately whether adult competence standards have been met.

While the poor predictive power and large errors of measurement of currently available formulas constitute the most serious barriers to using them in building a measure of reading effectiveness, other practical considerations also make these formulas unsuitable. No single formula is applicable across the entire range of difficulty to be covered by the reading effectiveness measure. Moreover, even within their applicable ranges, formulas are more accurate over some

ranges of difficulty than others (Chall, 1958). This is mainly a function of the difficulty range of the criterion against which the formula was originally validated, but may also reflect erroneous statistical assumptions in the construction of the formulas (Bormuth, 1966). Beyond the center of the range for which each formula was built, derived readability scores (arrived at by extrapolation or by an adjustment to the value yielded by the equation) tend to be so approximate as to serve no useful function for RRI's purposes (cf. Dale-Chall 1948).

Although it has been accepted practice in publishing and education to use different readability formulas for different segments of the difficulty range, this expedient cannot be used for the proposed reading effectiveness measure. The various formulas are not sufficiently alike to warrant treating them as though their scores reflected one continuous scale. The various formulas share neither predictors, criteria, nor computational procedures. Since growth in reading achievement on the proposed reading effectiveness measure is to be measured on a continuous scale, extending from the primary grades to adult levels, it is essential that all materials be rated with a formula based on a common set of predictors and criteria, and that the formula be equally sensitive across the entire difficulty range. Because existing formulas do not meet these requirements, there seems to be no alternative but to build a new readability measure.

This review of the shortcomings of existing formulas has been limited to technical problems with the formulas themselves. We have found that they do not predict enough criterion variance, that they have large errors of measurement, and that none is applicable over a sufficiently wide range of readability. Therefore we have not considered it necessary to question the validity of the criteria which existing formulas were built to predict. In the next section, however, dealing with selection of a criterion for a new formula, the reader will see that questions of criterion validity could legitimately be raised.

B. A New Readability Formula

In constructing a new readability formula, the two most important decisions that will need to be made concern the selection of a criterion scale of difficulty against which the formula will initially be validated, and the selection of predictor variables to be included in the formula.

1. Criterion scale of readability. Readability research has focused principally on predictors and, at least until recently, investigators have exhibited little concern for the quality of the criterion that is predicted. Authors publishing readability formulas usually have not reported on the reliability of their criterion measures (presumably because this reliability is unknown), although it is an accepted principle of measurement that successful prediction requires a reliable criterion. Some authors barely describe their criteria, even though proper interpretation of readability scores depends on a precise understanding of the criterion used in building the formula.

The quality of the criter... is essential to the utility of the formula. However astute an investigator may be in selecting variables that he believes should be predictive of the difficulty of text, the ultimate ability or inability of his formula to predict accurately the difficulty

of new passages will be a function of the validity and reliability of the criterion against which the formula is originally built. Since the criterion serves as the basis for accepting or rejecting potential variables for inclusion in the formula and for determining their weights, the better the criterion, the better those decisions are likely to be.

Three major kinds of criterion scales of readability can be identified in the research literature. These are: sets of passages scaled in terms of concurrent norm-referenced reading achievement test scores; publishers' grade level designation for books; and passages scaled for readers' ability to correctly guess deleted words. Each of these procedures for defining criterion scales will be discussed in turn.

a. Passages scaled against norm-referenced test scores. This type of criterion has been used more often than any other in the construction of readability formulas. Of this type, the most widely used set of criterion passages are the McCall-Crabbs grades test lessons in reading (1926), which were scaled in the following way. Students in grades three through six read 390 passages and answered multiple choice questions about them (seven through twelve questions per passage). The same students also took a standardized reading test. The grade placement for each passage was arbitrarily defined as the average reading grade level of students who correctly answered 75% of the questions for that passage.

The McCall-Crabbs test lessons (or other similarly graded passages) are unsuitable as a criterion scale for developing the required readability formula. There are two principal criticisms that invalidate this type of approach. One problem is that the grade levels assigned to the passages may not directly reflect the true difficulty levels of the material, since the scale values assigned depend on students' answers to multiple choice questions about the passages. Because it is a relatively simple matter to alter the comprehension scores that students can earn on a passage by changing either the type of question asked or the response options, the difficulty of passages measured this way resides as much in the questions asked as in the text itself. It is virtually impossible to demonstrate that the test items (questions and response options) are of comparable difficulty across passages. Noncomparability of item difficulty over passages would obviously make some passages easier (or harder) than they actually are, i.e., than they would be if item difficulty were controlled across all passages.[3]

---

[3] This type of imprecision in the criterion may partially account for the relatively poor prediction of criterion scores obtained from most readability formulas, noted above.

Second, the McCall-Crabbs passage scale values are based on relations between the percentage of questions answered correctly by students on that passage, and students' reading ability as measured by scores on a norm-referenced reading test. However, standardized reading test scores are themselves not directly interpretable with respect to what children can read. For one thing, norm-referenced scores are themselves based on multiple-choice comprehension questions whose difficulty is deliberately varied across passages. Thus norm-referenced scores reflect *something* other than just students' ability to comprehend passages of increasing difficulty. Using norm-referenced scores in an essentially circular way to define the "difficulty" level of other test passages seems likely to result in a criterion scale whose values do not correspond with precision to true differences in the inherent difficulty of reading materials. Thus the norm-referenced methodology underlying the McCall-Crabbs passages, or any other scales similarly constructed, makes them inappropriate to serve as a criterion of readability for the RRI reading effectiveness measure.

b. Publishers' ratings of books. Some authors of readability formulas (e.g., Spache, 1953) have used the grade level designations given to textbooks by publishers as the criterion scale of difficulty against which to validate their formulas. Since major textbook publishers control the

sentence length and vocabulary content of their books, especially in the early elementary grades, formulas that include sentence and vocabulary variables (as almost all do) would be expected to be good predictors of publishers' assigned grade levels. Whether formulas built to predict publishers' ratings of books are truly good predictors of readability is another matter.

At present, there is insufficient evidence to justify the assumption that publishers systematically increase the reading difficulty level of their instructional materials over grades. While publishers may control the number and familiarity of words and the lengths of sentences in their books to some extent during the early grades, there is evidence that they do not agree concerning which words to teach in which grades (Stauffer, 1966). Furthermore, Fry (1968) has reported that the readability of instructional materials changes more in some grades than in others (in an apparently random pattern). There is also reason to believe that the type and amount of control that publishers exercise over readability change as students get older (Spache, 1953).

Furthermore, even if publishers were to attempt systematically to control readability (e.g., by using existing formulas), it is not certain that the resulting materials would in fact be scaled for comprehensibility as intended. Attempts to alter the readability of passages by changing the

values of various formula components (e.g., simplifying vocabulary, shortening sentences) have not had any consistent effect on readers' measured ability to comprehend the altered passages (Klare, 1963). The failure to affect comprehensibility can probably not be attributed to differences between the difficulty of questions asked about original and altered passages, because questions were usually held constant over both versions of the passages. A more likely explanation is that current readability formulas fail to include enough of the important variables that affect the difficulty of text.

In short, publishers' grade level designations for books are not a suitable criterion for building the RRI readability formula because there is insufficient evidence that these designations are based on known difficulty characteristics of the materials. A formula is needed that is firmly anchored to properties of the reading materials, rather than one built merely to reflect what publishers believe children ought to learn in different grades.

c. Predicting deleted words: The cloze procedure. In recent years the cloze technique[4] (Taylor, 1953) has attracted attention as a new means for defining criterion scales of readability. In the cloze technique, words are

[4] The term cloze is derived from the concept of closure as used in Gestalt psychology. Closure refers to the human tendency to complete a familiar but not quite finished pattern (for example, to see a broken circle as whole by mentally closing up the gaps).

randomly or periodically deleted from text, and subjects are asked to guess the missing words. The cloze score is simply the percent of deleted words that are restored correctly. If several passages are deleted in a comparable way and presented to a group of readers for restoration, the passages can be ranked for readability according to their relative cloze scores: the higher the score, the more readable the passage.

The cloze technique represents a practical application to written English of research results that have confirmed speakers' ability to utilize the redundancies of language to extract information from garbled or incomplete messages. Language is characterized by rules that limit how elements (letters and words) may be combined, and by recurrent patterns that make some elements more probable in certain contexts than others. Because of these regularities, an element that is yet to come in a message is in some degree constrained by the elements that have preceded it.

For example, in English text, the letter "q" almost always signals that the letter "u" will follow. On the level of words rather than letters, the incomplete sentence "The man felt very ____" provides a great deal of information about the next word to occur. A user of English anticipates such words as "happy," "tired," or "weak." He would be surprised if the next word were "chimney," "there,"

or "drink." To the extent that what is to come in a sequence of words or letters is constrained to some extent by what has preceded it, the appearance of a particular word or letter is predictable to some degree from the context, and the sequence therefore possesses some redundancy.

The redundancy of English has been estimated to be 60-75% (Shannon, 1951; Garner, 1962). It is believed that this redundancy increases the likelihood that a message will be correctly received by slowing down the rate of information transmission, and by providing safeguards against the occurrence of communication failures due to accent, handwriting, noise, and ambiguities inherent in the language itself.

Since Shannon's (1948) initial applications of information theory to the study of language, a considerable body of evidence has been amassed indicating that users of English have learned to employ the redundancies of the language (Garner, 1962). Knowledge of these redundancies is demonstrated by users' ability to replace missing elements in a message, both at the level of letters (Chapanis, 1954; Miller & Friedman, 1957) and at the level of words (Aborn, Rubenstein & Sterling, 1959; Aborn & Rubenstein, 1958; Shepard, 1962).

More redundant sequences of words, by definition, should be easier to predict than less redundant sequences.

Taylor (1954) reasoned that the cloze technique, which tests subjects' ability to predict deleted words, should therefore provide a measure of redundancy. Taylor's (1954) research confirmed that cloze scores measure the redundancy present in text. He found that the cloze scores for deleted words had a rank order correlation of $r = .87$ with the estimated redundancy of those words in context.[5] Thus, cloze scores can be taken as a good estimate of the relative redundancy of language units in a passage of English text.

The degree of redundancy present in a passage of text and the readability of that text are related. The presence of redundancy reduces the amount of information transmitted in a message of fixed length. Therefore more redundant messages should be easier to comprehend--that is, more readable--than less redundant messages. Since cloze

---

[5] A computation of the redundancy of words in passages of English text by direct statistical analysis of sequences is presently unfeasible because of the size of the English lexicon and the difficulty of determining the distributional uncertainty of the words that occur in it. Instead, the redundancy of words can be estimated by assuming that subjects' predictions of words at a given location in a sequence provide good estimates of the probabilities governing the occurrence of the predicted words in that location. Subjects are presented with samples of English text that are $(n-1)$ words long and are asked to predict the nth word. The redundancy, R, of a predicted word is estimated by computing $\left[\frac{H_{max} - H}{H_{max}}\right]$, where H is the uncertainty (measured in bits) of

scores provide an estimate of the redundancy present in a passage of text, and since redundancy is related to readability, it follows that cloze technique should be able to measure the readability of text.

Studies comparing cloze scores with more traditional measures of readability comfirm that cloze scores do measure the readability of passages. Taylor (1953) found that the readability rank-ordering of three passages by both the Flesch (1948) and Dale-Chall (1948) formulas was reproduced by cloze scores. In two studies, Bormuth (1962, 1968a) found rank order correlations greater than $\underline{r} = .90$ between the ranking of passages by cloze tests and the ranking of passages by multiple choice comprehension tests.

The procedure for using cloze tests to determine the readability of a set of passages is simple. Comparable deletions are made in each passage removing an equal

---

The observed distribution of predicted words and H_{max} is the possible uncertainty, which is obtained when all predicted words occur equally often. For example, if a large number of subjects guessed three different words for a deleted word at some location in a passage of text, the maximum amount of uncertainty would be obtained if each word was used an equal number of times, i.e., each word occurred one third of the time. In these circumstances, the probability of occurrence of each word, p, would be $p = .33$ and, since H is given by $-\sum(p \log_2 p)$, $H_{max} = 1.58$ bits. If, however, the proportion of times each of the words was guessed was .50, .25, and .25, then the observed uncertainty is $H = 1.50$ bits, and, therefore, $R = \frac{1.58-1.50}{1.58} = .05$, or 5 percent.

number of words randomly or periodically (e.g., every fifth, seventh, tenth word) from each passage, and by replacing each deleted word with a blank of standard length. Taylor (1953) has shown that random and periodic deletion patterns yield equivalent data.

Deletions are made without regard for the function or meaning of specific words. Deleting only certain classes of words (e.g., only substantive words) is rejected because specified words or kinds of words may not occur equally often in different materials. Differences between passages in terms of the number of words occurring in different classes may itself be a readability factor, and its effect can be measured only by a method that operates independently of the number of words of different classes occurring in a passage.

The proportion of correct restorations per passage is the cloze score for that passage. The higher the cloze score, the more readable the passage.[6] Only exact

---

[6] The present discussion describes how cloze tests are used to measure the relative readability of several different passages. This is done by presenting several passages to some reader(s) and comparing the cloze scores earned by the several passages. Another use of cloze tests is possible. They can be used as a measure of students' reading comprehension. In the latter case, several students take a cloze test over the same passage(s) and readers' scores are compared.

Several studies have shown moderately high correlations between cloze scores and scores on standardized tests of reading (Taylor, 1957; Ruddell, 1965; Bormuth, 1965). However,

restorations (and obvious misspellings of exact restorations) are counted as correct, since Taylor (1953) has shown that the readability scores of passages are not affected by allowing credit for synonyms or by allowing partial credit for words that maintain the general meaning of the sentence.

Cloze scores have several important advantages over comprehension questions and over publishers' assigned grade levels as a criterion of readability. They are highly reliable (Taylor, 1953; MacGinite, 1971), whereas the reliability of other criteria is generally unknown. Between-passage differences in cloze scores are directly attributable to differences between the comprehensibility of the text of the passages, since estimates of passage difficulty are not affected by characteristics of the test items, e.g., by wording, response options, type of questions asked, etc. Variability in passage difficulty that could result from using one set of deletions rather than another set is easily controlled by using several different deleted versions (for example, some subjects restore every fifth word beginning

---

cloze scores should not be interpreted uncritically as a measure of comprehension. Salzinger, Portnoy & Feldman (1962) have shown that it is possible to correctly restore words to passages that are semantic nonsense, so long as the short-term contextual constraints of English are present. Chapanis (1954) found that the extent to which subjects successfully predict deleted units depends on their level of language skill. Thus it appears that language skill, per se, plays some role in determining cloze scores.

with the first word, others restore every fifth word beginning with the second word, etc.).

Finally, cloze scores have one other important characteristic that is not shared by other criterion scales of readability: they are known to be related to learning. Studies by Bormuth (1968b) and by Coleman & Miller (1968) have shown that the amount of information acquired from studying a passage is a function of subjects' original cloze scores on that passage. Since the RRI readability formula will be used to construct tests designed to measure the extent to which students have learned to read, it is desirable that the criterion scale of readability used to build the readability formula have a demonstrated relationship to learning.[7]

Recently, a criterion scale of readability covering a wide range of reading difficulty has been built with the cloze procedure (Miller & Coleman, 1967). Using college students as subjects, Miller and Coleman computed two cloze scores for each of 36 passages. One score (bilateral cloze) was based on the proportion of correct restorations made when subjects saw words on both sides of the deletions and the other score (unilateral cloze) was based on

---

[7] The relationship between learning (measured by information gain) and passage difficulty (measured by the cloze technique) is reviewed more thoroughly in Chapter V.

correct restorations when subjects saw only the words preceding the deletion, with all subsequent words masked out. Using these two sets of scores, which correlated r = .93 with each other, Miller and Coleman ranked the 36 passages from very easy to very difficult.

The validity of the Miller-Coleman scale has been demonstrated in a study by Aquino (1969). She found correlations above r = .90 between the Miller-Coleman scale values and the two independent procedures for ranking the same passages. These validating procedures were word-for-word recall of the passages, and judges' rank-ordering of passage difficulty. The fact that Aquino's subjects were drawn from a different population than the subjects used by Miller and Coleman lends increased weight to these findings regarding the validity of the scale.

An indirect test of the validity of the Miller-Coleman scale was provided by Szalay (1965). He used four readability formulas that had been developed using the Miller-Coleman scale as a criterion to predict the cloze scores subjects would earn on a new set of passages. Correlations between the actual and predicted cloze scores ranged from r = .83 to r = .89. The Miller-Coleman scale appears to be valid, since readability formulas based on it can be cross-validated at high levels of correlation.

In view of the advantages of the cloze procedure over norm-referenced scales and publishers' ratings, and in view of the validity data presented above, it appears that the Miller-Coleman scale is the best available criterion for the development of a readability formula for use in constructing the RRI reading effectiveness measure.

2. The need for a formula to predict readability. In view of the preceding discussion concerning the utility of the cloze procedure as a means for scaling the readability of passages, an explanation is in order concerning why the cloze procedure can be used only to develop a criterion for building a readability formula, rather than as a procedure for directly scaling adult-level reading materials and instructional materials used in the schools. In other words, why not directly apply the cloze procedure to scale the passages whose readability must be determined?

The reason that this cannot be done is practical rather than theoretical. It is true that direct scaling of passages by readers is feasible when a reasonably limited number of passages is to be rated. However, the quantity of material which may have to be rated for the reading effectiveness measure is so large that any direct scaling approach is, in effect, ruled out. A more practical alternative is to develop a formula for predicting the cloze scores that passages would earn if direct scaling were carried out. With

such a formula, the readability of any passage can be readily estimated without recourse to direct scaling by readers.

3. Variables that predict readability. Development of a formula for predicting readability requires that the structural and stylistic variables that discriminate between easier and harder passages be identified, and that the weighted combination of those variables capable of predicting criterion cloze scores with the greatest degree of accuracy be determined. The readability literature provides a good basis for at least a first attempt at selecting predictor variables.

Over the last 50 years, a great many variables have been proposed as possible indicators of reading difficulty. Factor analysis of those characteristics that have been the best predictors of reading difficulty has identified two major factors: vocabulary difficulty and sentence complexity (Brinton & Danielson, 1958; Stolurow & Newman, 1959). Of the two factors, vocabulary difficulty has been consistently the more important predictor. Thus it is not surprising that Klare's (1963) review shows that over half the formulas built to date include some type of vocabulary measure, while about one-third employ some measure of sentence complexity.

3.1 Measures of vocabulary difficulty.[8] Many measures of vocabulary difficulty have been used to predict readability.

---

[8] Numerous experimental results from the readability literature will be cited to support the discussion in this and the following section of the report. These results appear in the

Basically, vocabulary variables that have been used fall into three main classes: difficulty of vocabulary defined by the presence or absence of words appearing on a list of "easy" words; difficulty of vocabulary estimated by word length; and difficulty of vocabulary defined by some semantic property of the words, such as abstractness.

a. "Easy" words. In the first category, vocabulary difficulty is measured by comparing each word in the passage against a list of supposedly easy words. Each word in the passage is classified as easy or difficult according to whether or not it appears on the list, and either the number or proportion of hard (or easy) words is calculated for the whole passage. The widely used formulas developed by Lorge (1948), Dale & Chall (1948), and Spache (1953), employ word lists in this way.

The two lists most widely used are the Dale list of 769 words (Dale, 1931) and the Dale list of 3000

---

form of correlations between each of a number of predictor variables and some criterion scale of readability. In reading these results, the reader should bear in mind that the algebraic sign of the correlation (i.e., whether the correlation is positive or negative) does not affect the strength of the relationship between the variable and the criterion of readability. The strength of that relationship depends only on the magnitude of the correlation (i.e., the absolute size). Whether the correlation coefficient (r) is positive or negative depends on two factors.

First, the sign of the correlation depends on whether a higher value of the variable is related to more readable text or to less readable text. In some cases (e.g., proportion

words (Dale & Chall, 1948). The Dale list of 769 words originally contained those words that appear in both the International Kindergarten Union List (1928) and in the first 1000 words of the Thorndike-Lorge Teachers word book (Thorndike & Lorge, 1944). This list was updated by Stone (1956), who replaced 173 words with an equivalent number of words appearing more often in contemporary, primary grade, reading textbooks. The Dale 3000-word list contains approximately 3000 words "known" by at least 80% of fourth graders. The list was compiled by simply presenting lists of words to fourth graders and asking them to indicate by check mark which words they knew. When 80% of the students tested indicated that they knew a word, that word was included on the "easy" list.

---

of easy words), as the value of a predictor variable increases, readability increases (text gets easier). In other cases (e.g., proportion of hard words), as the value of the predictor variable increases, readability goes down (text gets harder).

The second factor affecting the sign of the correlation is the criterion scale of readability used in the research being reported. When the McCall-Crabbs (or similar) scale is the criterion, or when publishers' grade ratings of books are the criterion, higher scale values indicate less readable (harder) text. However, when cloze scores are used as the criterion of readability, higher scale values indicate more readable (easier) text. Therefore, if a variable correlates positively with McCall-Crabbs scores or publishers' grade level assignment of books, it must correlate negatively with cloze scores.

Estimates of vocabulary difficulty based on the presence or absence of a passage's words on an "easy" word list correlate rather well with criterion scales of readability. Lorge (1948), Dale & Chall (1948), and MacGinitie & Tretiak (1969) found correlations of r = .51, r = .68, and r = .63, respectively, between the proportion of words not on Dale lists and the grade levels of the McCall-Crabbs test lessons. Using publishers' assigned grade levels for textbooks as his criterion, Spache (1953) found a correlation of r = .68 between the proportion of words not on the Dale 769-word list and his criterion.

Using cloze scores as a criterion, Bormuth (1966) obtained correlations of r = .68 and r = .64 for the proportion of passage vocabulary appearing on the Dale 769-word list and the Dale 3000-word list, respectively, while MacGinitie & Tretiak (1969) found correlations of r = -.51 for the proportion of words not on the Dale 769-word list. (The smaller correlation coefficient of the latter study may be due to the fact that only one of five possible deletion sets was used to compute cloze scores.) The highest correlations with criterion scores yet reported for vocabulary difficulty based on word lists are reported by Coleman (1971), who found correlations of r = -.91 between Miller-Coleman scores and the ratio of words not on the Dale 3000-word list. It is likely that the difference between the magnitude of the

correlations reported by Bormuth and those reported by Coleman, both of whom used criteria based on cloze scores, is attributable to differences in the range of difficulty represented in the criterion scales. The Bormuth passages had an approximate readability range of grade 4.0 to grade 8.0, whereas the Miller-Coleman passages cover a range which appears to be at least twice as wide.

Recently, a more sophisticated procedure for classifying the difficulty of vocabulary on the basis of word familiarity has been proposed. Elley (1969) suggests using word frequency rather than simply presence or absence on a list of "easy" words. He argues that, since correlations which depend on a two-unit scale (e.g., presence or absence) are usually lower than those based on a graduated scale, a more refined measure of word familiarity (such as relative frequency of occurrence) should turn out to be an improved predictor of readability. He further proposes that only nouns be counted, on the ground that they are the least predictable elements in a passage and are, therefore, most critical to the understanding of a communication.

Elley computed the mean noun frequency value for 58 passages, using frequencies calculated from counts of words used by children. Across five validity studies in which judges' ratings of passage difficulty were correlated with the readability ratings based on the noun frequency counts,

the average correlation was r = .90 (range r = .85 to .95). The noun frequency count was a more powerful predictor of difficulty judgments than were all of 11 other predictors, which included two intact readability formulas and several major variables used in well-known readability formulas. Therefore, Elley's work suggests that graduated ratings of word familiarity may predict criterion scores more accurately than simple binary classification of words as easy or hard.

b. Word length. Word length has been used in some formulas as an index of vocabulary difficulty since, on the average, longer words tend to be less familiar (and hence, more difficult) than shorter words. Formulas using a word-length factor to measure vocabulary difficulty have included such characteristics as the number of syllables per 100 words, the proportion of monosyllabic and polysyllabic words, and average word length in letters and syllables. Dale & Tyler (1934) and Gray & Leary (1935) found that the percentage of one-syllable words correlated r = .38 and r = .43, respectively, with a criterion comprehension test. Later studies have shown higher correlations between word length measures and criterion scores. Flesch (1948) reported a correlation of r = .66 between average word length in syllables and McCall-Crabbs scale values. Bormuth (1966) found that average word length in syllables correlated r = -.89 with criterion cloze scores and that the corresponding correlation

for word length in letters was r = -.68. Coleman (1971) has reported correlations of r = .88 and r = -.90 between Miller-Coleman scale scores and, respectively, the number of one syllable words and the number of letters per word. Again, the larger correlations found by Coleman than by Bormuth are probably attributable to the greater difficulty range in Coleman's criterion scale.

c. Semantic word factors. The third type of estimate of vocabulary load requires some judgment concerning the semantic properties of the language in a passage. This approach is based on the assumption that, on the average, abstract words are harder to read and comprehend than concrete words. Therefore, counts have been made of many types of words presumed to discriminate between passages on an abstract-concrete continuum, including image-bearing words, sensory words, technical words, concrete ideas, abstract ideas, localisms, simple word labels, nouns of abstraction, finite verbs, definite words, realistic or specific words, references of an energetic, forceful, or vivid nature, formal versus popular words, definite articles, time nouns, and interjections. Because of imprecise definitions of the above variables, it is hard to know just how much overlap there is among them.

Although correlations as high as r = .68 have been reported between at least one "abstraction" variable (definite articles) and a readability criterion (Gillie, 1957),

formulas using a semantic approach to measuring vocabulary load have not been as successful, in general, in predicting criterion variance as have formulas using either word lists or word length (Klare, 1963). An apparent reason for their limited success is that there is little agreement as to how abstraction can be objectively defined.

3.2 Measures of sentence complexity. The second major factor affecting the readability of text is sentence complexity. Many different measures of sentence complexity have been tried in readability formulas. These may be grouped into measures of sentence length, prepositional phrase measures, and measures of syntax.

a. Sentence length. The most frequently used measure of sentence complexity has been sentence length, i.e., the average number of words per sentence. The rationale for using sentence length as an indicator of difficulty is that longer sentences are, on the average, more complex than shorter ones. Correlations of r = .47 (Lorge, 1948; Dale & Chall, 1948), r = .52 (Flesch, 1948), r = .57 (Coleman, 1971), and r = -.58 (Bormuth, 1966) have been reported between sentence length in words and various criteria of readability. Spache's (1953) reported correlation of r = .75 between mean sentence length and publishers' primary grade, textbook-level assignments is probably inflated by the fact that publishers control sentence length in primary grade texts.

Recent research by Bormuth (1966) suggests that average sentence length, measured by counting total syllables or letters, may prove to be even better predictors of reading difficulty than sentence length in words. Sentence length in syllables and in letters correlated $r = .70$ and $r = -.67$, respectively, with a cloze score criterion. The same research indicates that independent clause length may be a more powerful predictor of readability than any of these sentence length measures, since letters per independent clause correlated $r = -.81$ with cloze scores.

b. Prepositional phrases. Another measure of sentence complexity that has been used in readability formulas has been a count of prepositions or prepositional phrases. Correlations between prepositional phrase measures and readability criteria have been reported as $r = .35$ (Dale & Tyler, 1934; Gray & Leary, 1935), $r = .43$ (Lorge, 1948), and $r = -.41$ (Bormuth, 1966).

However, there is some question as to the true predictive value of prepositional phrase counts in readability formulas. The Dale-Chall formula, which differs from the Lorge formula chiefly in its lack of a prepositional phrase variable, is a better predictor of criterion readability than is the Lorge formula. Moreover, MacGinitie & Tretiak (1969) found that the relative contribution of prepositional phrases to reading difficulty varied drastically from one sample of

McCall-Crabbs criterion passages to another. They also found that, after sentence length and word difficulty were taken into account (predicting 80% of the variance in Miller-Coleman scale scores), the ratio of prepositional phrases added less than one tenth of one percent to the prediction of criterion scores.

c. Syntactic analyses. An early attempt to assess directly the syntactic complexity of passages was made by Vogel & Washburne (1928), who counted the number of simple versus compound and complex sentences. Apparently, the power of these variables was not sufficient to gain them widespread use.

In recent years, attempts have been made to develop predictor variables that would measure syntactic complexity with analytic procedures derived from theories of transformational grammar. One such predictor is word depth, which summarizes the complexity of a sentence. Word depth is theoretically related to the memory load imposed by sentence structure during generation of a sentence. Each word is assigned a "depth" as a function of how many structural characteristics of a sentence must be kept in mind at the time the word is produced. The greater the number of characteristics to be remembered, the greater the depth. Determination of word depth usually requires a diagram of the phrase (or constituent) structure of the sentence.

The relationship between word depth and reading difficulty has been found to be positive, though the significance of the relationship is not yet clear. Correlations as high as r = .78 have been reported between mean word depth and the comprehension difficulty of passages (Bormuth, 1964). However, word depth scores are highly correlated (r = .86) with sentence length (Bormuth, 1966), and both Bormuth (1966) and MacGinitie & Tretiak (1969) found that word depth is no better a predictor of readability than is mean sentence length.

Another predictor variable derived from transformational grammar is the ratio of kernels to sentences or words. Kernels are the simplest sentence units that are transformed to make more complex sentences. For example: We applauded his brilliant performance is built up from three kernels: He performed; He was brilliant; We appluaded him. Sentences that contain many kernels are syntactically more complex than sentences that contain only one kernel. Coleman (1971) found a correlation of r = -.77 between cloze criterion scores and an indirect estimate of the number of kernels.

Other indices of syntactic complexity derived from transformational grammar have been proposed, such as the number and type of transformations (Brown, 1967), or depth of subordination and deletions from deep to surface structure (Chomsky, 1971), but these measures have not yet been tested as predictors of criterion readability scales.

Finally, Ruddell (1965) found that the readability of passages for an audience of children is significantly related to the frequency with which syntactic structures in the passages appear in children's speech. However, Bormuth (1966) found a correlation of only $\underline{r} = .13$ between passage difficulty and the frequency of structures in children's speech. Perhaps the factor of congruence between speech patterns and written syntactic structures affects passage readability only for children.

C. Construction of the RRI Readability Formula

1. Reasons for building a new formula. Earlier in this report, it was established that the ability to measure the readability of text accurately is essential to the development of the RRI reading effectiveness measure, since the validity of the reading comprehension section of such a test depends on the ability to define precisely the difficulty of material that students can read. However, RRI's review of the readability literature, summarized earlier in this chapter, led to the conclusion that existing formulas for scaling readability are not suitable for use in building the RRI effectiveness measure. Available formulas were built to predict criteria of questionable validity. In addition, they only account for about half of the variability in the criterion scales that they were built to predict. Their use results in relatively large errors of measurement, and they are limited in the range of difficulty to which they are applicable. Thus a better means for scaling text is needed than present formulas provide.

The review of criterion scales of readability led to the conclusion that the cloze procedure (where subjects guess words that have been deleted from text) is the most reliable and valid way currently available for scaling readability. However, the very large quantities of text that probably will require scaling during the course of the proposed work on

reading effectiveness make it impractical to use the cloze procedure to scale directly the readability of all the materials.

The practical alternative suggested is to build a formula to predict readability, using cloze scores of selected passages as the criterion scale during construction. It is proposed that the 36 passages of the Miller-Coleman (1967) scale be used initially for this purpose. As noted earlier, this scale is probably the best available criterion of readability. Unfortunately, because the Miller-Coleman bilateral cloze scores are based on relatively few subjects, these scores may not be as stable a criterion as the RRI formula will require. Therefore, RRI proposes to administer cloze tests on the 36 Miller-Coleman passages to an appropriately large sample of readers to establish highly stable scale values.

Once the improved Miller-Coleman scale values have been determined, the following general strategy will be used to build the formula. First, several variables that should be predictive of passage difficulty will be selected, and the values of these variables calculated for each passage. Second, the correlations of these variables with each other and with the criterion measure of reading difficulty will be calculated. Third, using this information, several different algebraic formulas that reproduce the improved Miller-Coleman

scores will be generated. Fourth, the best of the alternative formulas will be identified by comparing their ability to predict accurately and reliably the values of the Miller-Coleman criterion scale after it has been expanded to include many more passages. Fifth, the validity of the formula will be verified using a new set of passages.

2. Constraints in the construction of the formula. The volume of material that probably will need to be scaled for readability in building the RRI effectiveness measure is so large that any human processing of the raw text, such as hand counting of any predictor variables, must be ruled out for practical reasons. RRI's choice of predictor variables is thus restricted to those that can be calculated on a computer.

The requirement for machine countable predictors rules out, at least for the present, predictor variables such as word depth or transformations from deep to surfact structure, since these require hand analysis by a linguist. For the same reason, RRI cannot include stylistic variables of the type proposed by Chomsky (1971), such as figures of speech, unusual choices of vocabulary, unusual word orders, and unusual sentence constructions.

While using a computer imposes constraints such as these, it also provides an opportunity to consider many more variables than could practicably be counted by hand. Since

data reported by Bormuth (1966) suggest that the ability to evaluate new variables increases the chances of developing a powerful formula, the use of a computer provides a decided advantage.

However, in view of computer costs, RRI will probably confine its investigations to predictor variables whose calculation involves the simplest possible computer processes consistent with achieving adequate predictive power. While the calculation of predictor variables during the initial stages of the construction of the formula would be relatively inexpensive regardless of the complexity of the computer processes involved (since only a limited number of passages need to be scaled), computer costs are bound to mount when analyzing the several corpora (see Chapters IV and V) that will have to be examined over the course of the proposed test construction effort. Therefore, there are advantages to keeping the variables in the formula as simple as possible.

3. Stages in the construction of the formula. The construction of the formula will proceed through several stages.

3.1 The generation of algebraic formula(s). In the initial stage, what is believed to be a good set of predictor variables will be selected and the value of each variable in each of the 36 passages will be determined. After calculating the correlations of the predictor variables with each other and with the criterion, one or more regression formula(s) can

be composed to predict the improved Miller-Coleman scores. Since many variables are available to be included in the regression equation, it is likely that several formulas can be constructed that will exactly predict the criterion scale scores. During this first stage of constructing the formula, issues of economy on the computer will be ignored, since the maximum possible information about candidate predictors must be obtained.

a. The selection of predictor variables to be tested. Based on evidence of their predictive power in previous research, the following measures of vocabulary difficulty will be evaluated in building a first-order formula: presence or absence of a word on an "easy" word list, word length in letters, and word length in syllables. As soon as it is feasible to do so (i.e., once word probabilities are established from an analysis of a corpus of English materials), the frequency of occurrence of words will be tested as a predictor variable. Sentence measures that will be tried because of their demonstrated predictive power in previous work will include sentence length in words, sentence length in syllables,[9] and sentence length in letters.

---

[9] The number of syllables can be closely approximated by a count of the number of vowels, plus the letter "y" (Coke & Rothkopf, 1970).

In addition to evaluating variables that have proved to be useful in previous readability research, a number of other candidate variables that intuitively appear promising will be evaluated. The number of commas, and other intra-sentence punctuation marks, will be counted, since these should signal sentence complexity. The proportion of words starting with letter combinations that frequently signal relational words also will be counted, such as those starting with wh and th (excluding "the"). Because many words beginning with wh and th are common function words, they tend to be both short and on all lists of "easy" words. This fact could lead us to underestimate the difficulty of some passages, since the wh and th words may signal greater syntactic complexity than other words on "easy" lists of comparable length.

Stylistic variables have not usually been included in readability formulas because of the subjectivity and difficulty of rating them, even though it has been noted by writers in the field (e.g., Klare, 1963; Bormuth, 1966) that the inability of formulas to take stylistic variables into account reduces their predictive power. At least one mechanically calculable feature that should reflect elements of style will be evaluated, namely, variability in sentence length.

b. The resolution of technical problems. In the initial stage of formula development, problems associated with a possible combination of linear and nonlinear predictors must be resolved. Virtually every existing readability formula has assumed a linear relationship between all predictor variables and readability; yet, as shown in Bormuth's (1966) work and elsewhere, this assumption is probably not true. Thus linear correlation models may be unsuitable for determining the relationship of predictor variables to each other and to the criterion.

Problems associated with establishing the algebraic form of the formula also must be analyzed and solved. Virtually all readability formulas add the weighted values of predictor variables. However, it seems intuitively plausible that their product (or perhaps some weighted geometric mean) may prove to be a better predictor.[10]

3.2 Comparing alternative formulas. Assuming that the first stage (3.1) yields several formulas, all of which exactly predict the improved Miller-Coleman scores, they will

---

[10] At the time of writing (July, 1973), we have completed the first few steps towards building the formula. The 36 Miller-Coleman passages have been keypunched, and the resulting deck of some 500 cards has been thoroughly checked. A program has been written which can read the cards, separate the entries into individual words, and perform a variety of counting and averaging operations. Thus, as soon as the improved Miller-Coleman scores have been calculated, RRI will be in a position to proceed to the actual construction of the formula.

need to be compared to determine which one should be selected as the RRI formula. To make this comparison, new passages will be added to enlarge the original criterion scale, and the formulas that predict the expanded scale most accurately will be identified.

a. An expanded criterion. There are two reasons for evaluating the various formulas in terms of their ability to predict an expanded criterion scale. First, the 36 Miller-Coleman passages may not adequately represent the range or the intervals of difficulty of adult-level English text. Second, the observed scale values of the variables in the 36 passages may not be typical of a larger random sample of English text, i.e., the language in the 36 passages may, for some reason, be atypical. The addition of more data points increases the chances that the criterion will be truly representative of the readability of English text. Hence a formula that can predict such a criterion scale should also be able to predict the readability of almost any passage of English text selected at random.

To expand the criterion scale, cloze tests will be prepared for each of a large number of randomly selected passages of text. These tests will be administered to a

large number of readers[11] to establish stable cloze scores for the passages.[12] Once scaled for readability, these new passages will be added to the 36 Miller-Coleman passages to create an expanded criterion scale. In enlarging the criterion scale, RRI will include pairs of passages with duplicate cloze scores and single passages that duplicate the cloze scores of passages on the original scale, in order that formula reliability can be studied (see below).

b. Comparing the formulas. The ability of the various formulas to predict accurately and reliably the cloze scores of the expanded criterion scale will be compared. Since passages ranging from very easy to very difficult will be scaled in building the reading effectiveness measure, the ability of formulas to predict accurately over the entire readability range will be compared. Since the reading effectiveness test to be taken by any one student will cover a relatively small segment of the readability scale, the ability of various formulas to predict accurately within a narrow range of difficulty, i.e., the ability of the formulas to

---

[11] Highly competent readers should be used to ensure that an adequate spread in the cloze scores of the most difficult passages is obtained.

[12] Since as many as 200 passages may be included in the expanded criterion scale, it would be unreasonable to expect any one subject to take a cloze test for every passage. Therefore, statistical designs that allow for some set of n passages to be assigned to each subject will be required.

discriminate between passages whose readability difference is small, will also be compared. Finally, since there will be passages on the expanded scale that have duplicate cloze scores, it will be possible to compare the reliability of the various passages, i.e., the ability of the formulas to assign matching readability scores to passages that have matching cloze scores. After analyzing the results of all these comparisons, unsatisfactory formulas will be discarded.

3.3 Conducting trade-off analyses. Even after unsatisfactory formulas have been discarded, it is quite possible that several formulas will remain, all of which predict the expanded criterion scores with perfect or near perfect accuracy. Up until this point in the development of the formula, issues of cost have not been actively considered (apart from the imposition of constraints in the selection of predictor variables, discussed earlier). To decide among these alternative satisfactory formulas, trade-off analyses will be carried out in which the power of each formula is evaluated against the computer costs associated with using it in calculations. These trade-off analyses should enable the most cost-effective formula to be identified.

In carrying out these analyses, particular interest will be focussed on examining the trade-offs associated with

choosing a formula that has a large number of variables, especially if any of those variables are costly to calculate.[13]

In principle, the predictive power of any formula should be increased by incorporating into it more and more variables that correlate with the criterion. However, previous research in readability (e.g., Bormuth, 1966; Farr, Jenkins & Paterson, 1951) suggests that after a few of the most pertinent variables have been used to compose a regression formula, the additional contribution to r^2 of more variables becomes very small. Therefore, in selecting the final formula, the contribution of each variable against its costs will be weighed carefully.

3.4 Verifying validity. Finally, after the most cost-effective formula has been selected, its ability to predict accurately the cloze scores of an entirely new set of passages must be verified. Since the formula will have satisfied rigorous criteria during the selection process, its ability to predict scores of new passages is reasonably assured. However, classic measurement theory requires cross-validation, hence RRI proposes to carry it out.

---

[13] In order to carry out these trade-off analyses, it may be necessary to solve some computer systems problems. Probably the largest problem will concern the practical use of word frequencies as a measure of vocabulary difficulty. To make use of word frequencies, ways will need to be found to reduce the computer time and costs currently associated with table look-up operations.

To carry out the cross-validation, a large number of subjects would be asked to take cloze tests over a new set of passages. To check the validity (i.e., the predictive accuracy) of the formula, cloze scores predicted by the formula for these new passages would be compared with the scores actually earned by the passages.

Chapter IV

Implementing the Design Concepts in the Construction of Reading Tests

The development of the readability formula and word-type distributions discussed in the preceding chapters will make possible the construction of tests to assess students' knowledge of words in various frequency of occurrence bands and students' ability to comprehend passages of text of various levels of difficulty. In this chapter, the procedures for building such tests will be outlined. The discussion will not cover the details of test construction, since the task of developing a "test blueprint" did not fall within the contract period covered by this report; rather, the discussion covers the major design concepts for building tests of reading achievement, noting differences between these concepts and current practices in norm-referenced test construction.

A. The Specification of a Corpus of English Text

The first step in the test construction plan is the specification of a corpus that can be used both to compute the frequencies of words that occur in adult English text and to scale the readability of adult reading materials. The accuracy of the readability and word frequency data will depend on the extent to which materials selected for the corpus are representative of the universe of written English that adults encounter. To be truly representative of this universe, the materials that make up the corpus must span the range of difficulty found in English text and must include all types of reading materials used by the adult population, in proportion to their extent of use.

To insure the representativeness of the corpus, two requirements must be met. First, because the universe of written English is very large, a domain (or subset) of materials must be identified that is comprehensive and that is amenable to systematic and objective sampling. Second, the domain must be systematically and objectively sampled to form the corpus. Thus, in a manner of speaking, the corpus emerges as a representative sample of a representative sample.

The need for objectivity and comprehensiveness in the construction of a corpus led RRI to examine the feasibility of using the contents of the Library of Congress as a domain

representative of the universe of written English. The Library of Congress is the largest general collection of adult reading materials available in this country. If text from each category in the Library's classification system were sampled in proportion to the size of the Library's holdings in that category, a representative corpus of English text should be obtained.

However, a preliminary examination of the classification system used by the Library revealed that the system is probably not amenable to systematic statistical sampling. The highly complex classification system is enumerative rather than analytic, making it hard to determine what constitutes a category of materials. Moreover, the classification scheme differs somewhat in each major division of the Library (presumably to meet the particular needs of each division) and is constantly being extended. Consequently, the size of the holdings in different categories would be difficult to determine. When these problems were uncovered, it was decided that, because of the classification system used, the Library's collection should not be used to obtain a representative sample of materials.

A simpler and more satisfactory procedure for obtaining a representative corpus of adult materials has been devised. This procedure is based on the identification of the domain of all periodicals as representative of the universe of written English and as an appropriate domain from which the RRI corpus

can be built through systematic and objective sampling. This domain includes newspapers, magazines, journals, and other materials published at regular intervals.

Periodicals should reflect the full spectrum of activities and concerns of society, and therefore should be representative of adult reading materials with respect to content. Their text should vary widely enough in difficulty to be representative of the readability of written English. The availability of such reference works as the Reader's Guide to Periodic Literature makes it possible to employ unbiased and systematic sampling procedures, such as the use of random numbers to designate which pages and which entries per page should be sampled. Furthermore, since circulation figures are available for periodicals, text from different periodicals can be sampled in proportion to the size of their circulations. Thus, the domain of periodicals is amenable to systematic sampling procedures to define a representative, adult corpus.

Practical considerations also dictate the use of periodicals as a domain from which the corpus will be defined. Assuming that a representative sample is drawn from current issues, the corpus can be assembled easily and inexpensively. There should be no problem locating the materials, and no likelihood of encountering archaic language.

Thus, periodicals appear to be a promising domain from which an adult corpus could be formed. However, there are two possible objections that could be raised to the use of periodicals, both of which are amenable to empirical resolution. One objection is that periodicals do not cover as broad a range of subject areas as books. The validity of this objection can be tested by randomly sampling books from a reasonably comprehensive collection (e.g., a large library) and determining the extent to which the subject matter of the sampled books is or is not contained in periodicals. The second objection is that the vocabulary and readability of materials within a subject area may differ systematically between books and periodicals. This objection can be tested by drawing random samples from books and periodicals in any field of study, and comparing them.

B. Calculating the Readability Values of Text and the Familiarity of Word Types

When the adult corpus has been identified, assembled, and entered into a computer, the readability of and word type distributions in the materials that make up the corpus can be calculated. The readability of materials in the corpus will be determined using the formula described in Chapter III. It should not be necessary to analyze the materials in the corpus in their entirety to calculate their readability values. Rather, it should be possible to base readability data on sample passages from text. Further study will be required to determine the number and length of passages that must be analyzed per periodical to achieve various levels of reliability in the readability estimates. Since longer passages and larger numbers of passages usually give more stable estimates of readability than shorter and fewer ones, the levels of reliability desired need to be weighed against the costs of increasing the size of the sample. These sampling decisions must be made before the formal analysis and scaling begin.

To determine word type distributions, every word in the materials selected for the corpus will be tabulated. For this purpose, a series of lemmatization rules are required, defining when words are to be counted as same or different. As noted in Chapter II, the lemmatization problem has been attacked by

other investigators, and it is possible that adequate procedures already exist for carrying out word-type frequency counts. If existing lemmatization procedures can be applied to the analyses of materials selected for the RRI corpus, many months of work will be saved.

C. Trade-Off Analyses: Precision vs. Complexity of Information

Before tests can actually be built, decisions must be made concerning the range of readability values and the number of word familiarity bands to be measured. These decisions cannot be made arbitrarily; rather, they depend on the outcome of trade-off analyses that weigh the costs and benefits associated with attaining precise measurements against those associated with attaining complex information from the reading test. Test construction is constrained by the fact that increasing either the precision or complexity of information obtained from the test raises testing time and costs. When time or costs are fixed, increases in precision can be gained only at the expense of complexity, and vice-versa. The need for precision in measurement must therefore be carefully examined in relation to the need for complex information.

To be useful in assessing growth in reading achievement, test scores must be replicable within small error, i.e., scores must have high reliability. The need for a reliable measure in detecting changes can be illustrated by a simple example. Suppose the weight of a person before and after a two-week diet is to be compared to see if he has lost weight. If the scale used is reliable only within five pounds, an observed difference of two pounds cannot be regarded with confidence

as a real loss in weight. Because of the relatively low reliability of the measuring instrument (the scale, in this case) any two weighings could yield as great a weight difference as that observed over the two-week time interval. Thus, the fact that the scale is only reliable within five pounds defeats the purpose of the measurement task at hand, which requires the detection of smaller differences. This scale may be perfectly adequate for many purposes, but not for taking measurements where differences of less than five pounds need to be detected reliably.

Similarly, if scores on a reading test are not reliable within a sufficiently narrow range relative to the change to be detected, the measurement error may be too large to permit a firm conclusion that an observed difference represents a true, rather than a chance, change in test score.

If only gross changes are of interest, such as differences between the reading skill of a first grader and a twelfth grader, a test which detects large changes reliably would be adequate to the task. But since it is essential to detect growth over much shorter time periods, e.g., from the beginning to the end of a school year, the test must measure small changes reliably to allow differences of this small magnitude to be detected with confidence. The capability of measuring small changes reliably requires, by definition, a high precision of measurement.

In order to detect small differences, it is necessary to minimize the chance variability in test scores. Some variability in test scores obtained at different times almost always occurs. Part of this variability may be attributable to real changes in the skill being measured, i.e., the subject has become more capable during the time interval between measurements. However, there are factors other than a true change in skill that can account for differences in test performance. To increase the likelihood of detecting a true change in performance, these other sources of score variability must be minimized.

The principal source of chance score variability is the test itself. Each item on a test represents a single observation of behavior. Relative to the total number of observations that could be made (i.e., the universe of test items that could be written), the number of observations actually made on any test is small. Whenever few observations are made, there is the chance that unusual or atypical instances of behavior will be observed and, because the total number of observations is small, will have a marked effect on the total score. For example, one set of items can turn out to be much easier for a particular student than another set because, by chance, the

first set contains a few items whose subject matter is particularly well known to the student.[1]

The most common way to reduce variability in scores stemming from item sampling peculiarities is to increase the number of items. Longer tests yeild more stable estimates of performance because they are based on a greater number of observations and, therefore, represent a more adequate sample of performance. As more items are sampled, the chance factors associated with individual items have less influence on the total score, and the total score is therefore more stable.

While in theory there is no barrier to increasing test length (the more observations that are made, the better), test length in practice is constrained by several factors, especially by the available testing time and the hourly cost of keeping a student in school. With time and budget limited, the use of as many items as possible to increase accuracy in measuring one skill often conflicts with the desire to use the test to

---

[1] Item sampling factors are not the only source of score variability. Temporary conditions unrelated to the skills being measured can also affect scores. For example, the subject may feel ill at one testing time, and hence not perform as well as he otherwise might; the room may be overly stuffy at one testing session; and so on. The effects of such sources of unreliability are present to some extent in all cognitive test scores. While these sources of unreliability present real problems in measurement, their control is properly the concern of test administrators and test interpreters, rather than the concern of test constructors.

measure more than one skill, i.e., to obtain complex information from the test. Unfortunately, with limited available time, increases in the precision of measurement can be obtained only at the expense of loss in the complexity of information. This poses a real dilemma, since the need for accuracy may make it impossible to measure all the skills about which information is desired.

Let us assume that a total of 30 minutes is available for testing vocabulary. If all 30 minutes are devoted to testing students' knowledge of words in one frequency of occurrence band, a precise measure of students' word knowledge in that frequency band probably could be obtained. However, no information would be obtained about their word knowledge in other frequency bands. On the other hand, if all frequency bands are tested, the number of observations made in each band may be too small to yield sufficiently precise and reliable data for detecting growth.

The level of precision required in a test will depend on whether group growth or individual growth is to be detected. Group scores are more stable than individual scores because pooling data over N individuals results in a degree of precision in the group score equivalent to having N times the number of observations on a single individual. Consequently, a test designed to detect group growth can employ fewer test items

to measure any single skill to the required degree of precision, and hence can measure a larger number of skills during a testing session of given length, than a test designed to detect individual growth.

Another element in the compromise between precision of measurement and the number of different skills measured (complexity of information) depends on whether the test will be used to determine if a student has attained a specified minimum standard of reading competence, or whether it will be used to determine the student's precise level of reading ability. Less testing time is required to determine whether or not a student has reached some minimum standard of reading competence than is required to specify his level of reading ability.

To determine whether a student has met minimum standard of competence, precision is required only at one point, X, on the scale (i.e., is his ability equal to or greater than X), whereas to determine a student's level of reading ability, precision is required at two points, X and Y, on the scale (i.e., is his ability equal to or greater than X and equal to to or less than Y). Therefore, if it is decided to test only for attainment of minimum standards, the lower requirements for testing time should make it possible to measure a larger number of skills than if the student's precise level of achievement needs to be determined.

An important element that will be required in the trade-off analyses is information concerning the quantitative relation between test length and test reliability for tests of the type discussed in this report. While it is generally true that longer tests (in which many observations are made) are more reliable than shorter ones, it is important to note that the well-established procedures for calculating the reliability of norm-referenced tests, and for estimating the effects of increased test length on the reliability of norm-referenced tests, and for estimating the effects of increased test length on the reliability of norm-referenced tests, are not applicable to the proposed RRI reading effectiveness measure.

The inapplicability of norm-referenced test methodology to determining the reliability of the RRI tests is not surprising, in view of the different purposes of these types of tests and the different concepts of reliability that follow from these different purposes. The purpose of a norm-referenced test is to discriminate among different persons' performances on a particular set of test items, while the purpose of the RRI test is to measure different persons' performances relative to some set of standards. Thus the concept of reliability for norm-referenced tests is framed in terms of stable discrimination, i.e., reproducing (e.g., in two or more

test administrations) the same rank-ordering among the members of a group of subjects irrespective of the numerical values of their scores, while the concept of reliability for the RRI test must be framed in terms of stable performance relative to the standards, i.e., reproducing (e.g., in two or more test administrations) the same numerical score for each subject taking the test.

Norm-referenced tests calculate *reliability* by correlating two sets of scores (e.g., from two administrations of different versions of the test) for each of N persons. Since correlation is mostly sensitive to the ordinal relationships among the scores in the two sets, reliability in norm-referenced tests is largely determined by the replicability of the relative size of scores, regardless of how large or small the numberical values of those scores may be. As long as a test places subjects in the same high-to-low order with repeated measurements, reliability as measured by any correlation coefficient will be high.[2]

In the RRI reading effectiveness measure, however, reliability should be defined and calculated in terms of the

---

[2] Note that $r = .975$ when X and X^2 are correlated for the integers 1 to 20 (i.e., when the sets of correlated numbers are 1 and 1, 2 and 4, 3 and 9, 4 and 16, ... 20 and 400, etc.). Thus r is very high even though the values of the numbers within each pair of scores differ considerably from each other.

degree to which the numerical values of test scores are replicable. A test with a "norm-referenced" reliability coefficient of $r = .99$ would be unreliable for RRI's purposes if all scores on one test were uniformly higher or lower than those on another. Therefore, when the mathematical properties of the reading effectiveness scale have been determined, new measures of test reliability appropriate to the purpose of the RRI reading test need to be developed.[3]

---

[3] Until an empirical attempt is made to construct and use the proposed reading effectiveness measure, we will not have sufficient information to define an arbitrary zero. In addition, until data are available, determination of whether the scale formed by reading effectiveness scores is continuous over all intervals, or whether it meets the requirements of an interval scale cannot be made. These and other issues need to be investigated before a meaningful measure of reliability can be formulated. For these reasons, a formal discussion of the issue of overall test reliability has been deferred in favor of concentrating on identifying and reducing those factors which contribute noise to measurement.

D. A Plan for the Efficient Use of Testing Time

Whatever decisions emerge from the trade-off analyses concerning the test's level of precision and complexity of information, testing time should be used as efficiently as possible. One way to increase the precision of measurement obtainable within a fixed period of time is to use a branched testing strategy (Chronbach, 1970). The principle of branched testing is to locate rapidly the student's approximate level of achievement on the skills being tested, and then to assign to each student a test whose items are concentrated around that level. Location of the student's approximate achievement level can be accomplished either by first giving a short, broad-spectrum test or by using information concerning performance on a previous test. By concentrating testing right around the level of the student's achievement, the number of relevant observations is increased, and it is possible to obtain a more precise and reliable estimate of his current reading skill level.

In measuring paragraph comprehension, the advantage of using a branched testing strategy is clear, since little information can be gained by using testing time to have students read passages that are much too easy or much too difficult for them.

In measuring vocabulary, the application of a branched testing strategy is less obvious. Ideally we would like to be able to measure students' word knowledge in several frequency bands. However, a certain minimum level of reliability for word knowledge scores in any one band must be obtained, or it will be impossible to detect growth in that band. Therefore, it may be impossible to test in as many frequency bands as might be desired.

Branched testing strategy would call for concentrating testing in the band or bands where it is most important to detect growth. This band is likely to change as students get older. For example, with younger students it may be most important to detect growth in knowledge of common words, whereas with older students it may be most important to detect growth in knowledge of moderately rare words. Some empirical evidence will be needed concerning children's knowledge of words in various frequency bands at different ages before deciding how best to implement a branched strategy in testing vocabulary.

E. A Plan for the Construction of Non-Biased Tests

In order to provide unbiased estimates of students' reading achievement, the tests should meet certain criteria of fairness. First, item content should be unbiased with respect to programs. All students, regardless of the program of instruction they have received, should be equally prepared for the test items. Thus, there must be sufficient overlap between programs in the readability of materials used and in their vocabulary content to yield a set of words and a range of readability on which all students can be tested. If such overlap does not exist, the tests would be biased against students in some programs.

Evidence suggests that in the elementary grades there is relatively little overlap between major reading textbook series with respect to the grades in which particular vocabulary words are introduced. Stauffer (1966) analyzed the vocabulary introduced in each of the first three grades in seven basal reader series. He found (Table 2) that the number of new words common to all series in each grade was remarkably small. Analyses are therefore required to determine the earliest grade at which the major reading programs are sufficiently similar to one another in the readability and vocabulary of their instructional materials to make it feasible to build tests that will be unbiased with respect to programs.

Table 2

Vocabulary Words Introduced in

Seven Basal Reader Series[a]

Grade	Total number of new words introduced	Number of new words common to all series
1	570	117
2	1,289	13
3	2,155	7
Total	4,014	137

[a]Based on Stauffer, 1966.

Second, test items should not penalize some groups of children or give unfair advantage to others (as has been charged of some norm-referenced tests) by including content that is likely to be more familiar to some groups than others because of ethnic background, social class, or place of residence.

Third, all test items should be as objective as possible. While standardized, norm-referenced tests are called "objective," the term applies only to scoring procedures. It does not apply to item construction, which is subjective in both the selection of questions and the generation of response options. In the proposed reading effectiveness measure, the goal should be objective item construction as well as objective scoring procedures.

Fourth, reading and comprehension of the passages should be both necessary and sufficient to answer the test questions rrectly. In some standardized tests, it is possible to answer certain items correctly without having read a test passage, because the item deals with general information that the student may have, independent of the material provided. In others, it is possible to miss certain items, even after having read and understood the passage, because the answer does not appear in the material provided. It is inappropriate to draw any conclusions concerning a student's reading skill

unless we can be certain that he has both read the test passage and that the passage provides all the information needed to answer the questions asked.

F. A Proposed Item Format[4] for Passage Comprehension

With the above considerations in mind, it is proposed that items for the reading comprehension test be written in a quasi-cloze format. In this item format, a word is deleted from text and replaced with a blank of standard length. The student must choose the deleted word from among several options provided following the end of the passage. The rationale for this item format is similar to that governing the use of the true cloze procedure as a measure of comprehension: the better a student understands what he is reading, the better

---

[4] The following additional item formats may be used if a procedure can be worked out for inferring from a student's responses to a set of questions measuring discreet pieces of information that he comprehends an entire passage.

Identification of missing facts or ideas. In this item format the student is asked to identify which of several ideas, people, problems, etc. is not mentioned in the passage. The student selects his answer from among several options, all but one of which have been mentioned in the passage. This item type is designed to measure students' understanding of the facts presented in a passage. Wording is altered between text and response options so that simple word matching will not yield the correct answer.

Vocabulary meaning in context. Several standardized tests purport to measure word meanings in context, but examination of the actual items shows that often they are ordinary vocabulary items, with context having little or no effect on word meaning. The item type proposed here is one in which the student must understand the passage to select the correct meaning of the word in context, since all response options will be genuine meanings of the word being tested. The correct response will alternate randomly between the dominant and secondary meanings of a word. Since most common words have

he should be able to guess a word that has been deleted from text. Items of this type are used in several major standardized tests.[5]

While this item type bears some resemblance to items in a true cloze test, it differs from regular cloze items in a number of important respects. Most importantly, in the proposed format the subject chooses the answer from among several response options provided, rather than generating the missing word himself. In addition, many fewer words are deleted in the proposed format than in a true close test (e.g., five

---

several meanings, it should not be difficult to construct such items with the help of a good unabridged dictionary.

Question about facts in passages. Bormuth (1968) has suggested that the item writing process can be made objective by constructing questions that are interrogative grammatical transformations of the syntax of sentences appearing in the passage. To make items of this type, a word, phrase, or clause is deleted from a sentence and is replaced by a question marker, transforming the sentence into a question for which the correct answer is the element that was deleted. For example, when various transformations are applied to the sentence "John rode the horse at the farm," the following questions result: Who rode the horse at the farm? and What did John ride at the farm? and Where did John ride the horse?, etc. This procedure permits a kind of control of item difficulty across passages, since the number and form of transformations (hence, questions) can be specified in advance, and randomly assigned to passages. By using this procedure, subjective judgments of item writers concerning the suitability and comparability of questions can be avoided. This seems to be a simple and direct way to determine whether the reader understands facts that are explicitly stated in a passage.

[5] Quasi-cloze items are used in the current editions of the Gates-MacGinitie Reading Test and the Stanford Reading Achievement Test.

words per passage of 100 words or longer rather than periodic deletion of every fifth, seventh, tenth, etc. word).

Earlier (see footnote 6 Chapter III), we noted that cloze tests have one potentially serious shortcoming as measures of comprehension, namely that the redundancy of English facilitates correct restoration of words even when the meaning of a passage is not understood. To avoid this problem in using quasi-cloze items in the reading effectiveness measure, function words (articles, conjunctions, etc.) will not be deleted, since grammatical knowledge alone can lead to their correct replacement even when the content of the passage is not understood (MacGinitie, 1971). Instead, only content words (nouns, adjectives, verbs, adverbs) will be deleted. To reduce further the likelihood that the constraints of English will lead to a correct response even though the material is not understood, all response options within each item will be equated for part of speech, word frequency, and plausibility when inserted in the deleted space. With these constraints imposed, a correct response should occur only when (apart from guessing) the student comprehends what he has read.

G. Construct Validity

There are two principal questions to be answered in establishing the construct validity of the proposed reading comprehension test. First, do the test items adequately measure the construct "reading comprehension"? Second, do the scoring procedures result in reasonable inferences concerning a student's comprehension of what he has read?

1. Do items measure "reading comprehension"? As described above, only one type of question (quasi-cloze) will be used in the reading comprehension test. It is assumed that this type of item taps a general comprehension factor. No items are proposed at this time for testing specific comprehension subskills, such as recognizing facts, drawing inferences from what is said, getting the main idea, understanding the author's purpose, discerning mood, recognizing literary devices, etc.

The decision not to measure reading subskills is a deliberate one made for the following reasons. We saw in an earlier section of this chapter that, in a test of fixed length, complexity of information can be obtained only at the expense of precision of measurement. Therefore, a decision to measure multiple skills would have to be based on a judgment that such complexity is worth obtaining, even at the expense of precision in the measurement of any one skill.

However, available data make it appear doubtful that the measurement of many subskills is worthwhile at the cost of test precision.

A review of reading tests by Berg (1973) has uncovered more than 70 distinct reading skills that test publishers have tried to measure. This probably reflects a widespread belief among educators that reading involves multiple skills and abilities. However, a number of studies (e.g., Thurstone, 1946; Harris, 1948; Hunt, 1957; Bormuth, 1969) indicate that the separate evaluation of all these factors is not warranted. As discussed earlier, the evidence seems to support the existence of only two principal factors: knowledge of individual words and comprehension of connected text. The data suggest that, apart from the apparently distinct word knowledge factor, the variance in reading test scores can reasonably well be accounted for by a single general comprehension factor. Thus the measurement of multiple comprehension subskills does not appear to be warranted. There are those (e.g., Davis, 1968, Lennon, 1962) who argue that measurement of a few distinct comprehension subskills (e.g., drawing inferences) is justified; however, the demonstrated degree of independence of these subskills has thus far been relatively modest.

It should be noted that the lack of convincing evidence for the existence of separate comprehension subskills

does not necessarily mean that such subskills do not exist. It is possible, for example, that investigators' failure to obtain distinct comprehension subskills in analyses of test data occurred because test items were poorly written, and therefore did not adequately measure the skills they were supposed to assess. However, in the absence of clear evidence of important independent comprehension subskills, RRI believes it is preferable to restrict the test to one item type measuring general comprehension of what has been read. Measurement of one general comprehension factor, rather than many subordinate ones, should increase the precision of measurement possible in the test. Furthermore, the proposed item format, unlike items required to measure the various reading subskills and unlike other item types that tap general understanding of a passage, makes it possible to construct items objectively and, to some extent, mechanically.

If the empirical data are correct in suggesting that all reading comprehension subskills are highly interrelated, then use of one item type that taps a general comprehension factor should adequately measure reading comprehension. However, it will be necessary at some future time formally to establish that the test is valid, i.e., that it measures what is theoretically meant by "reading comprehension." One way to demonstrate the test's construct validity would be to

follow procedures of the kind used in previous studies investigating the number of independent factors associated with the construct "reading comprehension." Essentially, this would involve asking questions of the type we have proposed, as well as questions of the type suggested by those who believe that reading comprehension involves many independent (or nearly independent) subskills. Once students' responses to the various kinds of questions were collected, it would be a relatively straightforward task to determine whether the RRI items measure the same factors as are measured by other types of items.

2. Is test performance validly interpreted? Performance on the reading effectiveness measure should be interpreted in terms of a subject's ability to comprehend materials written at various levels of readability. In order to draw such an inference, we must first define the performance that will be considered acceptable as an indicator of comprehension. We must decide, in other words, how many questions (all, most, some, etc.) a student must answer correctly before we credit him with understanding the passage on which the questions are based.

a. Using expert opinion to define comprehension criterion. One way to determine the criterion of comprehension (passing score) for a passage is to ask experts to define it.

This procedure calls for an arbitrary decision as to the level of performance that will be considered acceptable. Since good models of comprehension do not yet exist, there is currently no rational or empirical basis for concluding that the passing score for a passage should be, say, 90% rather than 80%. Because the decision is an arbitrary one, no matter where the passing score is set there are bound to be those who will argue that a different criterion would be more valid.

The problem of having experts set a criterion of comprehension is compounded by the fact that the definition of satisfactory comprehension must be a function of the purpose for which something is being read. There are some cases (e.g., medicine labels) where 100% comprehension is vital, but there are many other cases (e.g., newspapers) where less than 100% comprehension may be adequate for functional purposes. Thus it is understandable that experts should disagree as to a single best criterion of comprehension. Until good models of comprehension are available, these disagreements are unlikely to be resolved.

b. Using probability statistics to define criterion of comprehension. A more productive strategy for defining a criterion of comprehension for the reading effectiveness measure may lie in a statistical approach to the problem. As the discussion in this section will show, probability statistics

so constrain the possible interpretation of test results that the statistical approach may well be the only viable one to use.

Assume that the RRI reading comprehension test will be composed of N passages, each at a different level of readability. Further assume that testing time will be limited to the amount of time normally devoted to a reading comprehension test (30-45 minutes) and that, as a practical matter, a total of about 30-45 questions can be asked.

Few questions could be asked about each of many passages, or many questions could be asked about each of few passages. Since each of these extremes has distinct advantages and disadvantages (with regard to the range of readability that may be tested and the number of observations of behavior that may be made at each level of readability tested), assume that a middle course is chosen, in which a moderate number of questions (four or five) is asked about a moderate number of passages (six to eight).

It is an inescapable fact of multiple choice testing that some items can be answered correctly by guessing. The probability of correctly guessing an item depends on the number of alternatives from which one may choose and the plausibility of each of those alternatives. In theory, if there are N equally probable response options, the probability of guessing an item correctly is 1/N. This is the assumption normally made by psychometricians, e.g., in correcting a test

score for guessing. However, as a practical matter, if an item is poorly written, one or more response options may not be regarded as plausible, and the probability of a correct guess may be greater than 1/N.

For purposes of the argument below, it will be assumed that all test items have four or five response options, so that the probability of guessing the answer correctly if all responses have an equal likelihood of being chosen is p = .25 or p = .20, respectively. It will also be assumed that, for extraneous reasons,[6] the probability of a correct guess could be as high as p = .50. Actually, in most cases, the probability of a correct guess probably will lie somewhere between the boundaries p = .20 (for five response alternatives) or p = .25 (for four response alternatives) and p = .50.

Table 3 shows the probability of correctly answering r out of N questions on the basis of chance alone. The p = .20 and p = .25 columns give the probabilities when standard psychometric assumptions are made concerning the equal

---

[6] The principal extraneous factor that could affect the probabilities of response options in the quasi-cloze items proposed is the comparative likelihood of occurrence of the various options in the sentence frames provided. As noted in Chapter III, some words are more likely than others to occur in a given sequence. Even if attempts are made to equate all response options for semantic plausibility, the sequential probabilities of English may affect the likelihood that various response options will be selected. Note that the possibility of unequal probabilities for the various response options is independent

Table 3

The Likelihood of Guessing Correctly Various Numbers of Questions when the Probability (p) of Guessing the Correct Response is:

p = .20, p = .25 and p = .50

Number of questions guessed correctly	Four Questions per Passage			Five Questions per Passage		
	p=.20	p=.25	p=.50	p=.20	p=.25	p=.50
5	--	--	--	.0003	.0010	.0312
4	.0016	.0039	.0625	.0064	.0146	.1563
3	.0256	.0469	.2500	.0512	.0879	.3125

probability of response options. The p = .50 column gives the probabilities for the hypothetical worst case described above.

Table 4 lists the cumulative probabilities of guessing at least four out of five, three out of four, and three out of five questions correctly when the probability of a correct guess is either p = .20, p = .25, or p = .50. Table 4 shows that the freedom to select a criterion of comprehension is severely constrained by the probabilities of correct guesses. If the criterion of comprehension is defined as at least three out of four questions correct, there is a risk that, in as many as 31% of the cases, a conclusion that students understood a passage will be drawn when, in fact, they were only guessing. If at least three correct responses are required when five questions are asked about a passage, a student who is only guessing could meet the comprehension criterion as often as 50% of the time.

If we could safely assume that when the test is constructed all responses will be equally probable and therefore the probability of a correct response will be p = .25 if four response options are provided and p = .20 if five are

---

of partial knowledge that the student may have concerning the correct answer. For purposes of this discussion, it is assumed that the student has no comprehension of the passage, but that he is a competent user of English and hence knows the sequential probabilities of the language.

Table 4

The Cumulative Probability of Guessing Correctly Various Numbers of Questions when the Probability (p) of Guessing the Correct Response is: p=.20, p=.25 and p=.50

Number of questions guessed correctly	Four Questions per Passage			Five Questions per Passage		
	p=.20	p=.25	p=.50	p=.20	p=.25	p=.50
All	.0016	.0039	.0625	.0003	.0010	.0312
At least 4/5	-----	-----	-----	.0067	.0156	.1875
At least 3/4	.0272	.0508	.3125	-----	-----	-----
At least 3/5	-----	-----	-----	.0579	.1035	.5000

provided. Table 4 indicates that a criterion of at least three questions correct out of four or a criterion of at least four questions correct out of five would be acceptable,[7] in that either criterion would reduce to six percent or less the chances of crediting a guesser with comprehension of a passage. However, since it might happen that the empirical probabilities will turn out to be other than 1/N, the conservative course is to select at least four questions correct out of five as the criterion of comprehension. The hypothetical worst case would then lead to an erroneous conclusion only about 19 out of 100 times rather than 31 out of 100 times, which could occur if $p = .50$ and the criterion of comprehension was set as at least three questions out of four correct.[8] Table 4 makes it clear that providing five rather than four response options (so that

---

[7] As a practical matter, it seems unreasonable to set perfect performance (four out of four or five out of five) as the criterion of comprehension, since students may, for a variety of reasons, miss an item even if they comprehend the passage: their attention may wander momentarily, they may misread the response option, and so forth.

[8] It might be possible to pretest all options for equal plausibility, and revise them as necessary until $p = 1/N$. However, part of the strategy for development of the reading effectiveness measure is to make item construction as mechanical as possible. With this objective in mind it is probably more cost effective to plan conservatively for the possible worst case of $p = .50$ than it is to pretest all response options.

p = .20 rather than p = .25) reduces only slightly the likelihood that a student will meet the comprehension criterion by chance, when compared with the much larger reduction that takes place by setting the criterion as at least four questions correct out of five rather than at least three questions correct out of four. Therefore it is probably not worthwhile to spend the increment of testing time required for a student to process five rather than four response options.

c. Determining the maximum difficulty level a student can comprehend. After each passage has been scored, the data must be examined to determine the highest level of textual difficulty that a student can comprehend. This means analyzing the pattern of results over all passages to find the point at which a student ceases to meet the criterion of comprehension on test passages.

Certain clear patterns of data will be easy to interpret, e.g., where a student has a passing score on all passages up to a certain readability level, but fails all passages that are more difficult. However, less clearcut patterns of data also may be expected, such as when a student has passing scores on all passages up to a certain readability level, then alternately passes and fails a number of passages prior to finally failing all subsequent passages. Procedures need to be developed for specifying the most difficult level

of material that a student can read in those cases where the data do not show an unambiguous break point.[9]

The validity of inferences drawn from the data can be put to a simple empirical test. Students can be given materials of readability levels that (according to the reading effectiveness measure results) they should and should not be able to comprehend. Comprehension of these materials would have to be demonstrated behaviorally, e.g., by following directions. If the effectiveness test data have been validly interpreted, the student should succeed with materials predicted (by the effectiveness test results) to be within his comprehension, but should fail on materials that are predicted to be too hard for him.

---

[9] Probability theory will again be useful in making these decisions. Questions may be asked, for example, concerning the probability of obtaining different sequences or patterns of passing and failing passages on a chance bases in moving from the least difficult to the most difficult passages.

H. A Strategy for Measuring Knowledge of Words

To ensure that the test actually measures knowledge of the selected test words, the responses for each item should be considerably (and uniformly) more familiar than the test word itself. It is necessary to avoid the type of item, found in some norm-referenced vocabulary tests, where one or more of the response options is less familiar than the test word. It is not possible from items of the latter type to draw valid inferences concerning a student's knowledge of words in the frequency band from which the test item has been sampled, since the student may know the test word but not its less familiar synonym.

To avoid this problem, the following strategy in constructing test items could be adopted. First, a narrow interval A on the word familiarity scale would be selected; the interval should be narrow enough so that all of the types belonging to it are approximately equally familiar. Then, a random sample of, say, 20 types belonging to A would be chosen as test words.

For each test word, four responses also would be selected, of which only one matches the test word in meaning. All of these response options should be drawn from interval B, which

contains words that are more familiar than the test words.[10] This is shown in Fig. 4. It is unlikely that all 20 test words will have synonyms that are more familiar than the test words and that fall in a familiarity band as narrow as the interval A. That is why the interval B is shown as wider than the interval A. The exact width required in the interval B in order to find a synonym for all test words chosen from the interval A is not yet known; what is important is that the two intervals must not overlap.

Items to test knowledge of word meanings can be written in two principal formats; the test word can either be presented in isolation or it can be presented in the context of a brief linguistic frame, such as a phrase. Each format has its own advantage.

Presenting the test word in isolation has the advantage that it does not require the student to read any connected text in order to answer the question; hence performance on items in this format should clearly reflect students' knowledge of individual word meanings. Presenting the test word

---

10 This strategy (of making response options more familiar than test words by selecting the options from a scale interval of greater familiarity) can be expected to break down when testing the commonest (highest frequency) words. By definition, the most familiar words will not have more familiar synonyms. Hence an alternative procedure, perhaps involving pictures, will need to be developed when testing very high frequency words, to assure that response options are not less familiar than test words.

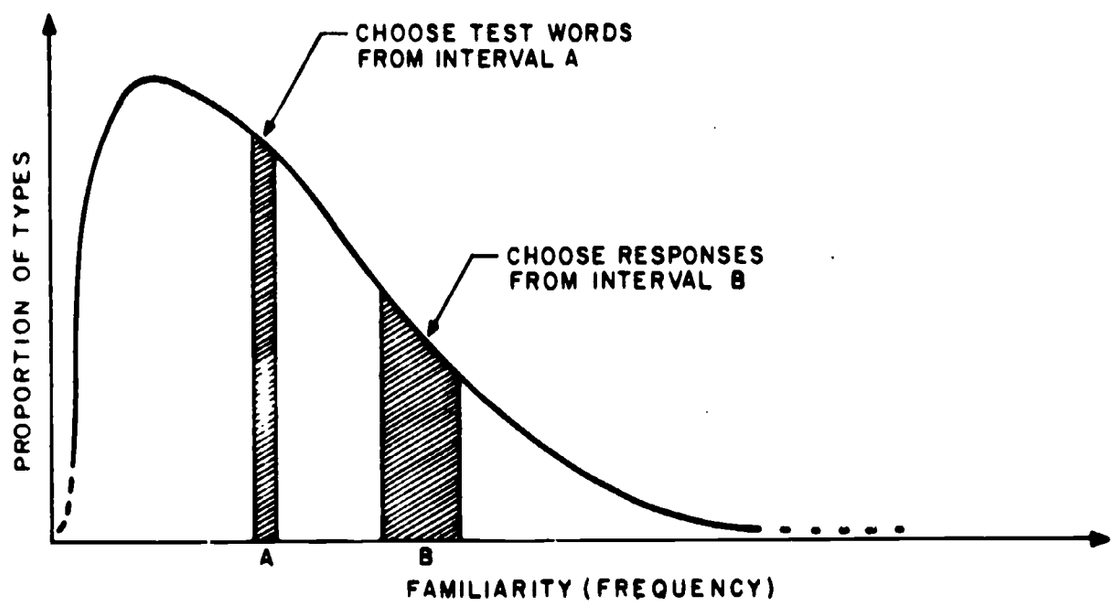


FIG. 4 THE SELECTION OF TEST-WORDS AND RESPONSES FROM DIFFERENT INTERVALS OF THE FAMILIARITY SCALE.

in a brief linguistic frame has the advantage of providing enough context to resolve ambiguities in the case of test words that have multiple meanings. Since multiple meanings are fairly common among familiar words, and since familiar words will be tested, it would be useful to have a format that could reduce uncertainty for the student concerning which of the meanings of a test word is intended in the test item.

A final decision on item format for measuring word knowledge has not yet been made.

I. A Plan for Computer-Assisted Test Construction

If a computer can be used to generate tests, then it will be practical to produce branched tests at costs that will be sufficiently low to be acceptable to school systems. Some computer assistance in the test construction process does appear to be feasible. In the comprehension section of the reading tests, a computer can help to generate alternative response options for the words that have been deleted from text. A computer can clearly be used to supply lists of words of comparable frequency to the deleted word. It is even possible that the list could be limited to words of the same form class as the deleted word. Item writers will need to make the final selection of response options, since options within an item should be matched for grammatic and semantic plausibility in the blank space (i.e., plausibility if a sentence stood by itself), and fulfillment of that criterion will require human judgment.

In the vocabulary section of the test, a computer can be used both to supply lists of words to be tested and to provide possible response options. Once a decision is made concerning the familiarity bands in which knowledge of words is to be tested, a computer can print out lists of words falling in the specified frequency bands. From these lists, test items can be selected. Once the synonym (correct option) for each

test item has been chosen (it is not known at this time whether this is best done by hand or computer), a computer can provide lists of words of comparable frequency. It may also be possible to restrict these lists to the same form class as the correct option. From these lists, selection of the required number of options will be made (or at least checked) by hand, since it is unlikely that a computer can be programmed at this time to do an adequate job with problems of multiple word meanings or connotations that could make an item ambiguous.

J. A Plan for Administering and Interpreting the Tests

Tests of reading comprehension and word knowledge built following the procedures described in this chapter can be administered periodically to assess students' reading achievement. At this stage the tests can be used only for limited purposes since, as yet, standards of reading competence have not been set. Although there is no basis, until standards are set, for judging whether students' achievement is "satisfactory," the tests can provide detailed information concerning what students know. By comparing students' performance over time, growth toward adult levels of reading comprehension and word knowledge can be measured.

The tests would be graduated in difficulty. Passages to test reading comprehension would gradually increase in difficulty from the lowest level of readability to the level of the most difficult materials found in the corpus. Vocabulary tests for the youngest students would start by testing knowledge of the highest frequency words, and tests for older students would gradually add words in the moderate and low frequency bands.

Scores on the comprehension test would indicate the level of readability of materials which a student can read with understanding. The scores could be interpreted for parents or teachers by illustrating the kinds of materials that a

student ought to be able to read, given his test performance. As a student gets older, his test performance could be related directly to adult tasks.[11] Scores on the word knowledge test would be interpreted by estimating the size of a student's vocabulary in various frequency bands, and comparing his word knowledge profile to word type distributions in adult materials.

---

11 While the purpose of these tests is to provide data that are directly interpretable with respect to students' reading capabilities, it is a simple matter to provide a norm-referenced interpretation for these test scores if such an interpretation is desired, by accumulating sufficient test data and reporting the distributions of scores per grade.

Chapter V

Application of the Design Concepts for Quantifying English Text in Setting and Monitoring Standards

In the previous chapter, the application of the readability formula and word frequency distributions to the construction of tests to assess students' current reading achievement and their progress toward attaining adult levels of reading skill was discussed. This chapter will show how the quantitative scaling of English text in terms of readability and word familiarity would facilitate the formal evaluation of reading effectiveness by providing relevant input data for setting standards of adult reading competence, and how this scaling can serve as a basis for measuring students' progress toward the attainment of those standards.

This chapter will also show how the readability and word familiarity measures could be used in analyses of instructional materials to determine whether or not the readability and vocabulary of those materials may account for students' failure to meet standards of adult reading competence. These analyses could result in recommendations for more rational readability and vocabulary demands on students which, if implemented, could lead to an increase in system effectiveness in reading.

A. Input Data for Setting Standards

Setting standards of adult reading competence requires value judgments to be made concerning the capabilities that adults need to acquire. Such value judgments are the responsibility of government. While scientists have no direct role to play in making "should" or "ought" statements concerning adult reading capabilities, they do have an important role to play in providing those empowered to set standards with the technical data upon which informed decisions should be based. Analysis of English text in terms of readability and word familiarity makes it possible to provide government with information in a form that enables standards to be set in precise, quantitative terms.

1. The quantitative display of adult English text. Suppose that the readability and word frequency characteristics of adult materials have been scaled, using periodicals to define the corpus of adult reading materials. The basic display of input data to be provided, for example, to the legislative branch of government should familiarize its members with the range and general meaning of scale values found in adult reading materials. This could be done by displaying the scale values of all the analyzed materials, clustered by difficulty and word frequency characteristics. Such a display would reveal the functional meaning of differences in scale values. For example, it might be seen that a cluster

at the low (easy) end of the scale includes such materials as the Readers' Digest, Daily News, pulp fiction magazines, etc. The upper end of the scale might include physics journals, literary criticism magazines, etc. This display should help anchor scale meanings for legislators in terms of materials which are familiar to them.

2. The difficulty and familiarity of the reading requirements imposed by New York State. Once the entire difficulty range has been arrayed, legislators may ask for the scale values of any *set(s) of materials* they consider relevant to setting standards. It can be anticipated, for example, that one set of materials whose scale values would constitute useful input data would be those materials that the State of New York expects its citizens to read. This would include such diverse materials published by state governmental agencies as tax forms, driver's license applications, official notices, advisory information on a variety of subjects, etc. Since publication of these materials implies that the government currently expects citizens to be able to read them, the readability and word frequency characteristics of these materials are highly relevant to the setting of realistic standards of adult reading competence.

The scaling of these materials would allow legislators and others in government to see how the reading tasks placed on citizens by the state compare with the full range

of adult materials. If any reading tasks considered essential should fall at the upper (difficult) end of the scale, government would have the choice either of setting standards at that high level, or of directing that an attempt be made to simplify the materials to some lower level that was specified as the standard.

However, the extent to which materials can be simplified is constrained. The heart of the problem in simplifying text is that vocabulary difficulty is the single most important factor affecting readability, and, in at least some cases (such as the insurance policies cited above, or tax forms, legal documents, etc.), the vocabulary required to convey essential concepts cannot be replaced by more common words. Thus, although some sentences may be made less complex and some simplification of vocabulary may be possible, it is unrealistic to expect that the difficulty of all essential or important reading materials could be reduced to elementary reading levels.[1]

---

[1] This fact is illustrated in a recent attempt by Pennsylvania's insurance commissioner to increase the readability of insurance policies. Using Flesch's prescriptions for producing more readable writing by manipulating such features as word length and sentence length, Blue Shield succeeded in changing the readability scores of a Medicare policy only slightly, from 26.8 to 35.0 (or about 8%), on a scale from 0 (very difficult) to 100 (very easy) (New York Times, July 8, 1973).

3. Scaling other pertinent materials. Since the materials issued by New York State may not cover all of the important reading tasks that adults need to carry out, government might request additional input data. Once word-type probabilities have been determined, and more is known about the range and relative frequency of different levels of readability in the corpus constituted from periodicals, any materials can be analyzed and located on the readability and word familiarity scales. Thus, if they wished, persons in government could receive data concerning the scale values of other reading tasks that have been considered important by presumed experts, such as the tasks in the Harris Survey, in the Educational Testing Service collection, or in the Adult Performance Level Study. At their request, they might also receive information on the comparative readability and word frequency characteristics of entry level materials in various fields, e.g., what levels of reading skill are required to read introductory texts in automotive mechanics, business administration, chemistry, etc.

4. Forecasting future reading requirements. In gathering input data for setting standards, government may also wish to receive information concerning the reading requirements that citizens will need to meet fifteen to twenty years in the future. A minimum forecast of twelve years would be required, since it takes that long for a

student to complete his schooling. The program of reading instruction that a student receives from the time he enters school should be geared to preparing him for the standards he will need to meet, at the time of high school graduation, in order to enter college, technical school, or an apprenticeship program, to get a job, or to function effectively in the adult world in general.

The need for forecasting is based on the assumption that adult reading requirements will be different in the future from what they are today. Linguists (e.g., Fries, 1962) have provided considerable information that language is constantly changing. However, such changes in language are slow to occur. A much more potent force producing short-term changes in the difficulty of adult reading materials results from the fact that our society is rapidly changing its requirements for citizenship and work.

An informal analysis of reading requirements suggests that the materials that people need to read in order to function effectively as adults have increased in quantity and in difficulty over the last few decades. In order to use the sophisticated machinery and the array of new products developed in recent years, adults must read more than they had to in earlier times. Consumer trends, such as increased use of credit agreements, checks, etc., also mean that the public must do more reading. Each year there are fewer jobs

that can be filled by persons with limited reading skills. These changes and others have resulted in a sharp increase in the reading demands on adults over the last generation. Since there is no reason to expect society's rapid rate of change to decline in the near future, the reading demands that will face the next generation of adults will probably differ from those facing current high school graduates. Therefore, to do an adequate job of educating students to be competent adult readers, we need to be able to forecast the reading skills that a student entering first grade now will need by the time he is a graduating twelfth grader.

The design concepts for quantifying language phenomena appear to make forecasting feasible. The general strategy for forecasting would be straightforward: in any field, materials would be sampled to establish the mean and variance of their readability at several historical points in time. A curve would be fit to the data, and from this curve extrapolations could be made to future points in time. This procedure could be followed for materials that citizens are required (by law) to read, and for materials adults need to read for their own well-being, as well as for materials in any occupational area. If government wanted to make the forecast data more sophisticated by weighting such factors as the future prospects for an occupation, or the anticipated

importance of various reading tasks, methods for quantifying such trends may have to be developed and weighting procedures would certainly need to be formulated.

5. Using input data to set standards. After examining all input information that it considers relevant, government can set terminal standards of reading comprehension and vocabulary for students in New York State, based on the readability and word frequency characteristics of reading tasks it considers essential or important. These would probably be system-wide rather than individual standards. That is, the standards would probably define a floor or minimum level of reading competence for any student being processed by the educational system; they would probably not purport to specify the standards that any particular individual should strive to achieve.

When standards are set on the basis of information provided by analyses of the readability and word familiarity of various reading materials, they lead to precise definitions of what the educational system must produce. For example, if New York State sets as a standard that all graduating students be able to read income tax forms, then schools must turn out students who can read materials of difficulty level X, and who know Y percent of common words, Z percent of rare words, etc.

6. Standards may be tentative until costs are known. Since the present state-of-the-art does not make it possible to calculate the costs associated with attaining various standards, government might wish to regard the standards it sets as tentative, pending information on how attainable they are, and at what cost. It may, however, be a long while before this information can be provided. In order to relate costs to the attainment of standards, the time (and therefore the costs) required for teachers to achieve various program objectives must be known. Furthermore, it is necessary to understand how achieving program objectives affects progress toward the attainment of standards (i.e., operationally, how achievement of program objectives affects performance on effectiveness measures). Finally, the consequences of achieving particular objectives must be known not only for short-term progress toward standards, but also for the long-term or ultimate attainment of standards.

Most of the complex information needed to accomplish this linking of time (costs), objectives, and standards is not presently available. For example, one prerequisite is programs that are fully described with respect to objectives and the procedures used for meeting those objectives. Since educational systems currently do not formulate precise program documentation, the linking of costs and standards does not appear likely in the near term.

B. Measuring and Displaying Effectiveness in Reading

With standards of adult reading competence defined, it becomes possible formally to evaluate whether educational systems (the state as a whole, districts, etc.) meet those standards. As noted in the preceding chapter, reading achievement can be measured even though standards have not been set, but a formal judgment cannot be made concerning the adequacy of achievement until there are standards against which to evaluate performance. Once minimum terminal objectives in reading have been specified, a criterion exists against which the success of educational systems can be evaluated.

1. Measuring attainment of standards. System effectiveness would be measured by administering tests of word knowledge and reading comprehension to high school seniors to determine whether or not their knowledge of words in various frequency bands and the readability of materials they can comprehend meet the levels specified by the standards.

2. Measuring progress toward standards. On the hypothesis that system effectiveness in reading were actually measured and the results reported, several questions can be anticipated in connection with interpreting and understanding the findings, particularly if system effectiveness were generally below minimum standards. For example, how do students move toward the standards over the grades? Does progress increase monotonically over the grades, or does it essentially level off after, say, grade six?

To answer such questions, it should be possible to measure students' progress toward adult standards by administering one or more tests in which readability and vocabulary vary from the very simplest levels up to the level of adult standards. The results of periodic testing could be related to the standards to determine whether students were progressing toward adult standards and to describe how far they had progressed.[2]

While student progress toward adult standards could be described, it would not be possible (at least in the near term) to evaluate the progress that was made, i.e., to say whether or not it was adequate. Since grade level standards of reading competence have not been established, there is presently no basis for judging that a particular level of reading skill in grade X is satisfactory or unsatisfactory with respect to reaching adult competence by the twelfth grade.

3. Measuring attainment of de facto grade standards. Although student progress cannot be formally evaluated relative to the ultimate attainment of adult standards, it is a

---

[2] The tests of graduated difficulty described here are similar to those described in the previous chapter, where the discussion concerned assessment procedures. The principal difference lies in the criterion used to interpret test data: in the present instance growth is measured against specific standards, whereas in the previous chapter growth was related to adult reading in general.

straightforward matter to determine whether or not students are meeting the de facto standards that are operative in grades 1-12. These de facto standards can be defined by employing the readability formula and knowledge or word familiarity (based on analysis of the corpus of adult periodicals) to characterize the instructional materials used in each grade.

In order to do this, the instructional materials used in each of the grades would first need to be sampled, thereby constituting several grade level corpora.[3,4] To determine grade level expectations (de facto standards) for readability, the formula would be applied to the corpora, and

---

[3] The grade level corpora complied by Carroll, Davies, and Richman (1971) would not be suitable because: (a) they lack information for several grades; (b) they may not accurately represent instructional materials used in New York State; and (c) the reported word frequencies do not take size of readership into account.

[4] In defining grade level corpora of instructional materials, it is probably unnecessary to analyze the readability and word type distributions of every book used in the schools. Books that are used for supplementary study and resource purposes in the classroom may be excluded, since many of them are used by few rather than all students, and thus do not reflect stable reading demands that schools are placing on students.

Library books should also be excluded, since it may be impossible to estimate accurately how many students use them or the grade(s) in which they are used.

On the other hand, limiting the corpus exclusively to reading textbooks would constitute too narrow a definition of the reading demands that are placed on students. To

the mean and standard deviation of the materials used in each grade would be calculated. To determine grade level expectations for word familiarity, the vocabulary used in each grade would be characterized in relation to the familiarity of words previously found in the analysis of adult materials. Once the de facto readability and word familiarity standards of each grade were defined in this manner, it would be possible to analyze students' performance on the reading effectiveness

---

function effectively, a student must be able to read, in addition to his formal reader, at least the contents of his math, social studies, science, and health books. Since use of these books is required of practically all students, their readability and vocabulary should be considered a legitimate part of the reading demands placed on students in the grades and programs in which they are used.

The information needed to construct the instructional corpus may be collected through a survey, asking which books are used as required texts in different schools in reading, math, science, health, and social studies, and in which grades those books are used. Since the number of different textbooks in these curriculum areas is limited relative to the number of schools, a survey of a stratified random sample of schools, rather than a survey of every school in New York State, should be sufficient to net all major texts.

After the corpus of instructional materials has been identified, readability and word type distributions would be calculated by grade, program, and content area. Weighting procedures would be used to take account of differences in the number of students using the materials.

measure to determine whether these grade level standards were being met.[5]

[5] It is apparent that the proposed measure of effectiveness in reading could also be used to determine whether standards other than those which may exist in a de facto sense at various grade levels are attained. If reading programs specified that they expected students to know vocabulary of a given familiarity or to be able to read at a given difficulty level, then whether or not a program was meeting its objectives in year Y, Y+1, Y+2, etc. could be determined by testing to see whether students had acquired the word knowledge and reading comprehension levels set as program objectives. Furthermore, with the appropriate statistical designs, it would also be possible to determine comparative program effectiveness, e.g., among programs A, B, C, and D, at the time of completion, and also to determine the comparative long-term payoff of the programs N years after completion.

C. Analyses of Effectiveness Data

In the event that administration of effectiveness measures revealed that students were not meeting standards, it would be important for senior managers at the State Education Department to identify the factors contributing to the less-than-desired levels of effectiveness. Logically, the current problem would have to be defined before alternatives for improving system performance could be formulated. The readability and word familiarity design concepts make possible a number of analyses that should contribute to an understanding of why standards are not being met, and should indicate possible courses of action to correct the problem. While the types of analyses outlined below would not be sufficient to identify all the factors contributing to students' failure to attain standards, they should make a distinct contribution to a good definition of some of the problems and alternative solutions.

1. Multivariate analyses of effectiveness in reading. Information concerning whether de facto grade standards were being met could be analyzed in conjunction with information concerning the pattern of progress toward adult standards over the grades in order to understand why any target group(s) or school system(s) were attaining a given level of effectiveness by the twelfth grade. Such a study would enable SED to isolate problems and consider alternative solutions.

For example, let us assume that a problem definitional study showed that all (black, white, and Spanish-speaking) middle class students attained the minimum adult standards by grade ten; all lower class (white, black, and Spanish-speaking) students did not attain minimum standards by grade twelve. Let us further assume that the de facto grade level standards increased monotonically from grades one through five and then leveled off so that the rate of growth in expected reading skill was less for grades six through twelve than for grades one through five; furthermore, by tenth grade, de facto standards had reached the level of adult standards. Continuing with our hypothetical example, let us assume that all middle class students met or exceeded grade level standards in all grades, but that, starting in grade three, all lower class students did not attain grade level standards. It then follows that it would be reasonable to inquire into the feasibility of reducing, for lower class students, the rate by which the standards increased in grades one through five, increasing the rate in grades six through twelve, and setting the present de facto grade ten standard as the expected value in grade twelve. Such changes might reasonably be considered in order to better insure that lower class students attained adult standards by graduation.

2. Analyses of instructional materials. The readability and word familiarity measures could be used in analyses

of instructional materials designed to locate possible problems in the readability and vocabulary learning demands placed on students. Once such problems were identified, categorical responses designed to alleviate them, and hence to increase effectiveness, could be suggested to senior managers of the State Education Department.

The analyses that are illustrated below do not, by any means, exhaust the analyses of instructional materials that might be carried out in looking for sources of failure to meet adult standards. For example, the question of the consequences of alternative methods of teaching reading on attaining standards is not touched upon. Rather, the illustrations are limited to analyses that are related to the readability and word familiarity measures.

a. Disorder (chaos) in instructional materials. To RRI's knowledge, no learning theorist or educator has ever found that chaos contributed to learning. Yet there is reason to believe that the vocabulary of instructional materials creates a chaotic set of inputs for students.

Evidence presented by Stauffer (1966) indicates that reading programs differ concerning which words are introduced in which grades. If there are major differences between the vocabulary in various sets of instructional materials, there should be difficulty in maintaining a logical sequence of instruction. Whenever a school changed books or whenever

a student changed schools, students would be apt to encounter a great number of words not previously learned or, alternatively, be asked to learn words already mastered. The problem of encountering a great number of new words could be expected to be most serious for those students who are most dependent on the schools for what they learn, presumably educationally disadvantaged students.

On the hypothesis that construction of grade-level corpora were undertaken, the data would be available with which to determine the degree of similarity across reading program materials from different publishers with respect to the vocabulary and readability introduced in each of the grades. If a display of the overlap of readability and of vocabulary across programs confirms that, in fact, there is appreciable chaos (i.e., little overlap), corrective action would be indicated. A logical step to correct the problem would be for the State Education Department to direct publishers that there be (at least) a minimum specified amount of overlap in vocabulary and readability across reading programs.

b. <u>The effect of the amount of learning expected per time unit (load) on cognitive development</u>. There is evidence to suggest that the vocabulary load that is being placed on students is not being carefully planned or controlled. First, there is evidence suggesting that publishers

do not coordinate vocabulary across curriculum areas. Stauffer's (1966) comparison of reading, arithmetic, health, and science books showed little overlap in each grade between the vocabulary words introduced in the books used in the different subject areas of the curriculum. For example, he found that while 2153 new words were introduced in seven reading series in the third grade, and 2150 new words were introduced in three arithmetic series in that grade, only 421 words were common to both lists. Moreover, many words appeared in textbooks in different subject areas which did not appear in any of the seven reading series at any grade level. Stauffer has estimated that even if a student somehow had the opportunity to learn the vocabulary of all seven reading series, he would learn only half the words he would encounter in his arithmetic books.

If these data are correct, the effective reading load on pupils is considerably heavier than the requirements made by formal reading programs (and, for that matter, heavier than the requirements made by typical, college-level, foreign language courses), since students must also learn much new vocabulary in the subject areas. Furthermore, since new vocabulary is seldom identified as such in textbooks, other than basal readers, there is a good chance that teachers will not formally teach these new words, thereby leaving the learning of much critical vocabulary entirely up to the student.

There is also evidence suggesting that the reading load presently placed on students may not be evenly distributed over grades. Stauffer's (1966) data show a geometric increase in the number of new reading vocabulary words introduced per grade, in grades one through three.

RRI believes that loading rate should have important effects on reading achievement since a related variable, massed versus distributed practice, is a classic learning variable of considerable significance. However, the actual relations between loading rate and learning to read are unknown, and would need to be clarified in a program of research before the load currently found in instructional materials could be evaluated.

Once the relationships between loading rate and learning to read were known, that knowledge could be used to evaluate current loading practices in terms of whether or not they promote efficient learning. For example, it could be that the vocabulary load of reading programs is about right for efficient learning but that, in some grades and for some types of learners, the added burden of vocabulary in the subject areas creates an overload.

A possible outcome of such analyses could be directives from the State Education Department to publishers governing the number of new words to be introduced per grade, and the coordination of vocabulary across curriculum areas.

c. The effects of difficulty on learning. The readability formula and word frequency distributions could be used in analyses designed to determine whether the difficulty level of instructional materials being used by students might be responsible for their failure to attain adult standards of reading competence.

It is well established that difficulty is a critical variable in learning. When materials are too easy for students, boredom results. When materials are too difficult, students may become frustrated, "tune out," and develop negative attitudes toward learning. In some cases, students may even develop negative attitudes toward themselves (e.g., "I am not capable of learning") or set unrealistically low achievement aspirations in order to reduce the impact of the failure experiences they are likely to have when asked to learn materials that are too difficult for them. In short, when materials are either too easy or too hard, students do not learn efficiently and the risk that educational processes will produce unwanted outcomes increases. It is for these reasons that controlling the difficulty of instructional materials has been a traditional concern of educators.

As long ago as 1917, Thorndike (1917) suggested that teachers not use materials for instructional purposes unless a child could correctly answer 75% of the comprehension questions asked of him about the materials after he had studied them. Many well known writers in the field of

education have since echoed Thorndike's suggestion in writing for teachers. The long history of interest in measuring readability is another reflection of educators' desire to present students with materials of a suitable level of difficulty.

A recent study by Bormuth (1968b) clearly indicates that the amount learned from studying instructional materials is a function of the initial difficulty of the materials for the student. After matching pairs of students for reading skill, Bormuth had one member of each pair take a cloze test over a passage to measure his comprehension of the passage, while the other member of the pair answered questions about the passage prior to and after reading it. He correlated cloze scores with information gain, which was defined as the increase in the number of questions answered correctly after reading the passage. Bormuth found (Fig. 5) that students gained very little information from studying difficult materials (cloze scores equal to and below 25%), and that studying easy materials (cloze scores equal to and higher than 37%) resulted in only very slight increases in information gain.

Bormuth's data indicate that there is a level of difficulty--neither too easy nor too hard--over which each increment in difficulty will result in a corresponding increment in information gain. As Fig. 5 shows, information gain for Bormuth's subjects was a monotonic function of difficulty

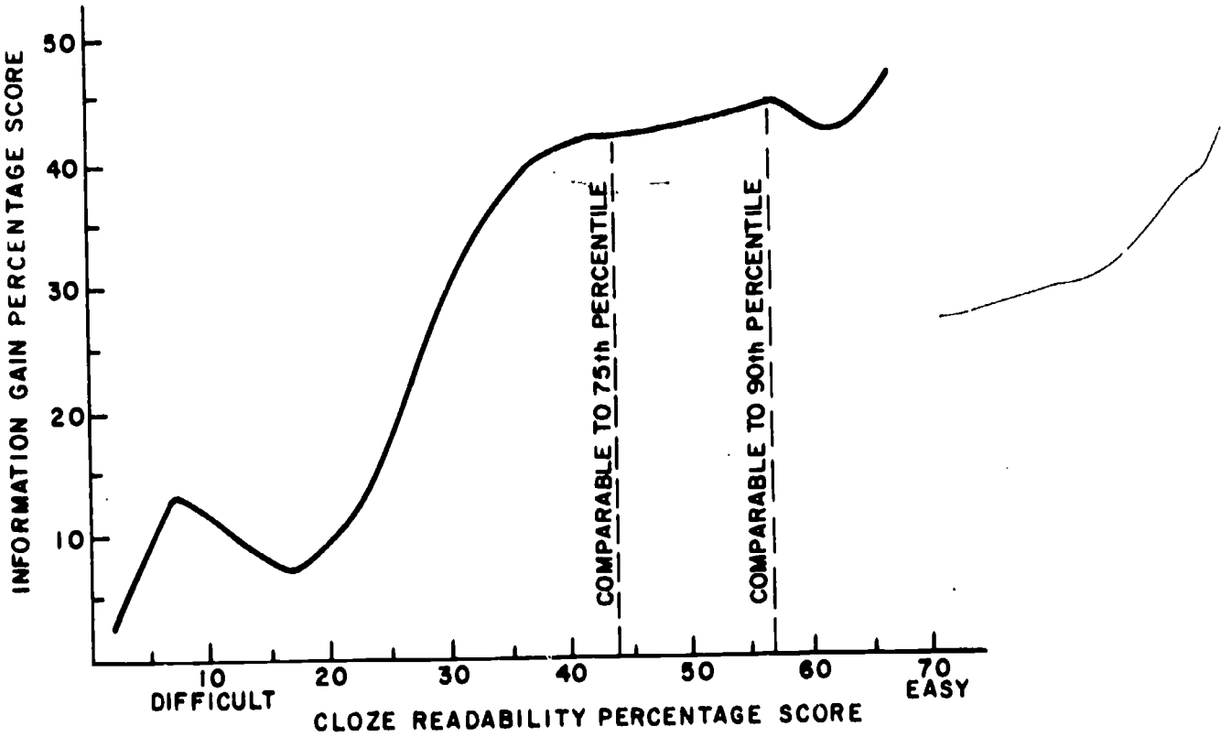


FIG. 5 EIGHT DEGREE POLYNOMIAL CURVE FITTED TO THE REGRESSION OF EACH PAIR'S INFORMATION GAIN SCORE ON ITS CLOZE READABILITY SCORE. (SLIGHTLY MODIFIED FROM BORMUTH, 1968b.)

only over the range bounded by cloze scores of 25 to 37 percent.[6]

As Bormuth notes, however, we cannot assume that the learning curve found for his college subjects would hold for all learners over all materials. Rather, the results in Fig. 5 should probably be taken as illustrative of a more complex set of relationships; it may be assumed that they represent only one of a family of curves that would result if the relationships between learning difficulty and information gain were studied for students of different ages, abilities, and backgrounds, as well as for different types of of materials and amounts of study time. Further research would be required to define this family of learning curves before the the appropriateness of the difficulty level of particular instructional materials for particular students could be evaluated.[7]

---

[6] It is interesting to note that Thorndike's (1917) suggestion that teachers use moderately easy materials (at least 75% comprehensible) was a good estimate (see Bormuth [1968b] for how the 75th and 90th comprehension percentiles were translated into cloze scores). It appears from Fig. 5, however, that, in terms of cloze scores, a more difficult average level of difficulty should govern the material presented to students.

[7] To facilitate later use of the data, difficulty in such research should probably be defined in terms of readability discrepancy (i.e., the difference between the readability of the new materials to be learned and the readability of materials the student can currently comprehend) rather than in terms of cloze scores.

Once these relationships between difficulty and learning were established, it would be possible to determine whether unsuitably difficult instructional materials (either too easy or too hard) were associated with students' failure to meet standards. On the hypothesis that effectiveness measures were regularly administered to students, the readability levels that students could comprehend would be known. The readability of instructional materials used by students could be scaled with the readability formula. It would then be a straightforward task to determine the difficulty of the materials for students, and to evaluate that difficulty in terms of the learning curves that had been established.

It might be that some students were being asked to use instructional materials whose readability was, according to the learning curves, much too difficult for efficient learning. In surveys conducted among a nationally representative sample of eleventh grade students in 1960 and 1970, 38% and 33% of those interviewed, respectively, reported that, at least half the time, "I read material over and over again without really understanding what I have read" (Flanagan and Jung, 1971).

In the elementary grades, the problem of excessively difficult instructional materials is likely to be most severe in subject areas of the curriculum, since teachers have

greater latitude to match basal readers to students' reading capabilities than they have to assign science, math, or social studies texts that are in accord with students' reading skills. Normally, subject area texts are determined on the basis of topics to be covered in a particular grade, and these topics are considered to be fixed, independent of students' reading abilities.

To promote optimum learning, the recommendation might be made that teachers take steps to provide all students with materials of suitable difficulty. Development of the readability formula and determination of the optimal difficulty levels for learning should provide practical tools to assist teachers in doing so. While traditional readability formulas have been viewed as a means for helping teachers to judge the suitability of materials, the absence of clear guidelines concerning how to employ them (i.e., what readability level should be chosen for instructional purposes for a child currently reading at level X) has effectively precluded their use as a means for routinely selecting instructional materials.

Application of the guidelines proposed by Thorndike (i.e., that materials are suitable if 75% of questions are answered correctly) is so cumbersome that it is highly unlikely that teachers have made use of these guidelines on any regular basis. Bormuth's (1968b) suggestion

that cloze tests be substituted for multiple choice questions in determining the suitability of materials hardly seems to make the teacher's job much easier. By contrast, an accurate readability formula (for scaling all materials used in schools) coupled with knowledge of optimum difficulty levels for learning should make it possible to supply teachers with the information they need to determine the suitability of new materials for learners on a routine basis.

However, there is a possibility that the problem of inappropriately difficult instructional materials may not be readily solved. Suppose that the readability of science materials used in the fifth grade were too hard for some fifth grade students. An obvious recommendation would be that more readable materials be provided. This might require that publishers be directed to produce materials covering essentially the same content, but at easier levels of readability. However, as noted earlier in this chapter, the vocabulary required to convey essential concepts limits the extent to which the readability of materials covering a particular subject matter can be simplified. Thus it may not be realistic to expect that all the needed instructional materials could be produced even if the effort were made.

If materials of appropriate readability (and suitable content) could not be provided to students, it would pose a clear problem for the efficient management of instruction.

Traditional solutions of educators to the problem presented by students who cannot read the materials designated for their grades have included remedial instruction (to try to bring students to the point where they could profitably use grade level materials), non-promotion, ability grouping, and reduced class size. However, since the effects of these and other managerial policies on learning are uncertain, a recommended course of action for the State Education Department's managers is not evident at this time. The problem is a complex one that will require study.

D. A Suitable Measure and a Means for Change

The approach to the measurement of effectiveness in reading that has been described in this report differs in two significant respects from other attempts to measure system performance in reading. First, the proposed approach should result in tests meeting all the functional specifications for an effectiveness measure outlined in Chapter I. The report shows how the design concepts for characterizing language in terms of readability and word familiarity would contribute to setting precise quantitative standards of adult reading competence, and how they would make it possible to build reliable, valid, and clearly interpretable measures to ascertain students' progress toward and attainment of those standards.

Second, the approach described in this report should make it possible to improve system effectiveness through a chain of interrelated steps involving many of the key actors in the educational process: teachers, educational managers, and publishers. For example, it has been shown that, if system effectiveness were below par, the readability and word familiarity measures could be used in analyses to identify possible problems in the learning demands placed on students. These analyses might lead to recommendations for changes in the vocabulary and readability content of instructional materials in different grades. Since the recommended changes

would rationalize learning demands, they should increase the likelihood that students will ultimately meet adult standards of competence in reading. The measures could be used to verify that the designated changes were implemented and to determine whether the anticipated changes in achievement occurred. If they did not, the process could be recycled.

The prospect of using measurement as one step in a chain to increase system effectiveness clearly sets the proposed effectiveness measure apart from other attempts to assess system performance in reading. Other assessment procedures, such as norm-referenced tests, seldom if ever result in the identification of problems or in recommendations for change, because they lack mechanisms for examining test results in relation to instructional materials. By contrast, the strategy for measuring effectiveness in reading that has been proposed in this report could quite possibly lead to an increase in system effectiveness when fully implemented.

References

1. Aborn, M. & Rubenstein, H. Perception of contextually dependent word probabilities. American Journal of Psychology, 1958, 71, 420-422.
2. Aborn, M., Rubenstein, H., & Sterling, T.D. Sources of contextual constraint upon words in sentences. Journal of Experimental Psychology, 1959, 57, 171-180.
3. Adult Performance Level Project Staff. The adult performance level study. Austin: University of Texas, 1973.
4. Aquino, M.R. The validity of the Miller-Coleman readability scale. Reading Research Quarterly, 1969, 4, 342-357.
5. Berg, P.C. Evaluating reading abilities. In W.H. MacGinitie (Ed.) Assessment problems in reading. Newark, Del.: International Reading Association, in press.
6. Bormuth, J.R. Cloze tests as measures of readability and comprehension ability. Unpublished doctoral dissertation, School of Education, Indiana University, 1962.
7. Bormuth, J.R. Mean word depth as a predictor of comprehension difficulty. California Journal of Educational Research, 1964, 15, 226-231.
8. Bormuth, J.R. Comparisons among cloze test scoring methods. In J.A. Figurel (Ed.) Reading and inquiry. Proceedings of the International Reading Association, 1965, 10, 283-286.

9. Bormuth, J.R. Readability: a new approach. Reading Research Quarterly, 1966, 1, 79-132.

10. Bormuth, J.R. Cloze test readability: criterion reference scores. Journal of Educational Measurement, 1968, 5, 189-196. (a)

11. Bormuth, J.R. Empirical determination of the instructional reading level. Reading and Realism, 1968, 13, 716-721. (b)

12. Bormuth, J.R. Factor validity of cloze tests as measures of reading comprehension ability. Reading Research Quarterly, 1969, 4, 358-365.

13. Brinton, J.E. & Danielson, W.A. A factor analysis of language elements affecting readability. Journalism Quarterly, 1958, 35, 420-426.

14. Brown, D.L. Transformational depth. Inglewood, California: Southwest Regional Laboratory for Educational Research and Development, 1967.

15. Carroll, J.B. Word frequency studies and the lognormal distribution. In E.M. Zale (Ed.), Proceedings of the conference on language and language behavior. New York: Appleton-Century-Crofts, 1968.

16. Carroll, J.B., Davies, P. & Richman, B. The American Heritage word frequency book. Boston: Houghton Mifflin and New York: American Heritage Publishing Co., 1971.

17. Chall, J.S. Readability: an appraisal of research and application. Columbus: Bureau of Educational Research, Ohio State University, 1958.

18. Chapanis, A. The reconstruction of abbreviated printed messages. Journal of Experimental Psychology, 1954, 48, 496-510.

19. Chomsky, C. Linguistic development in children from 6 to 10. Cambridge, Mass.: Radcliffe Institute, 1971.

20. Coke, E.U. & Rothkopf, E.Z. Note on a simple algorithm for a computer-produced reading ease score. Journal of Applied Psychology, 1970, 54, 208-210.

21. Coleman, E.B. Developing a technology of written instruction: some determiners of the complexity of prose. In E.Z. Rothkopf and P.E. Johnson (Eds.) Verbal learning research and the technology of written instruction. New York: Teachers College Press, 1971. Pp 155-204.

22. Coleman, E.B. & Miller, G.R. A measure of information gained during prose learning. Reading Research Quarterly, 1968, 3, 369-386.

23. Cronbach, L.J. Essentials of psychological testing (3rd ed.) New York: Harper and Row, 1970.

24. Dale, E. A comparison of two word lists. Educational Research Bulletin, 1931, 10, 484-489.

25. Dale, E. & Chall, J. A formula for predicting readability. Educational Research Bulletin, 1948, 27(1), 11-20, 28; 27(2), 11-28.

26. Dale, E. & Tyler, R.W. A study of the factors influencing the difficulty of reading materials for adults of limited reading ability. Library Quarterly, 1934, 4, 384-412.

27. Davis, F.B. Research in comprehension in reading. Reading Research Quarterly, 1968, 3, 499-544.

28. Educational Testing Service. ETS takes part in national right to read effort. ETS Developments, 1971, 18(3), 2.

29. Elley, W.B. The assessment of readability by noun frequency counts. Reading Research Quarterly, 1969, 4, 411-427.

30. Farr, J.N., Jenkins, J.J., & Paterson, D.G. Simplification of Flesch reading ease formula. Journal of Applied Psychology, 1951, 35, 333-337.

31. Finn, P.J. Syntactic and semantic complexity as criteria for scaling instructional materials. Paper prepared for the Bureau of School and Cultural Research, New York State Education Department, March, 1973.

32. Flanagan, J.C., & Jung, S.M. Progress in education: A sample survey (1960-1970). Palo Alto, Calif.: American Institutes for Research, 1971.

33. Flesch, R. A new readability yardstick. Journal of Applied Psychology, 1948, 32, 221-233.

34. Fries, C.C. Linguistics and reading. New York: Holt, Rinehart, and Winston, 1962.

35. Fry, E. A readability formula that saves time. Journal of Reading, 1968, 11, 513-516, 575-578.

36. Garner, W.R. Uncertainty and structure as psychological concepts. New York: John Wiley and Sons, 1962.

37. Gillie, P.J. A simplified formula for measuring abstraction in writing. Journal of Applied Psychology, 1957, 41, 214-217.

38. Gray, W.S. & Leary, B.E. What makes a book readable: an initial study. Chicago: University of Chicago Press, 1935.

39. Gunning, R. The technique of clear writing. New York: McGraw-Hill, 1952.

40. Harris, C. Measurement of comprehension of literature: the nature of comprehension. School Review, 1948, 61, 280-289, 332-342.

41. Harris, L. & Associates. The 1971 national reading difficulty index: a study of functional reading ability in the U.S. for the National Reading Center. New York: Harris and Associates, 1971.

42. Herdan, G. Type-token mathematics. The Hague: Mouton, 1960.

43. Hilliard, R.M. Massive attack on illiteracy. American Library Association Bulletin, 1963, 57, 1034-1038.

44. Horn, E. A basic writing vocabulary: 10,000 words most commonly used in writing. University of Iowa Monographs in Education, 1926, Ser.1, No.4.

45. Howes, D. A word count of spoken English. Journal of Verbal Learning and Verbal Behavior, 1966, 5, 572-604.

46. Howes, D.H. & Solomon, R.L. Visual duration threshold as a function of word probability. Journal of Experimental Psychology, 1951, 41, 401-410.

47. Hunt, L.C. Can we measure specific factors associated with reading comprehension? Journal of Educational Research, 1957, 51, 161-172.

48. International Kindergarten Union, Child Study Committee. A study of the vocabulary of children before entering first grade. Washington: International Kindergarten Union, 1928.

49. Klare, G.R. The measurement of readability. Ames, Iowa: Iowa State University Press, 1963.

50. Kucera, H. & Francis, W.N. Computational analysis of present day American English. Providence, R.I.: Brown University Press, 1967.

51. Lennon, R.T. What can be measured? The Reading Teacher, 1962, 15, 326-337.

52. Lorge, I. Predicting reading difficulty of selections for children. Elementary English Review, 1939, 16, 229-233.

53. Lorge, I. The Lorge and Flesch readability formulas: a correction. School and Society, 1948, 67, 141-142.

54. MacGinitie, W.H. Discussion of Professor Coleman's paper. In E.Z. Rothkopf and P.E. Johnson (Eds.) Verbal learning research and the technology of written instruction. New York: Teachers College Press, 1971. Pp 205-215.

55. MacGinitie, W.H. & Tretiak, R. Sentence depth measures as predictors of reading difficulty. Reading Research Quarterly, 1971, 6, 364-377.

56. Maginnis, G H. The readability graph and informal reading inventories. The Reading Teacher, 1969, 22, 516-518, 559.

57. McCall, W.A. & Crabbs, L.M. Standard test lessons in reading. New York: Teachers College, 1926.

58. McLaughlin, G.H. SMOG grading--a new readability formula. Journal of Reading, 1969, 12, 639-646.

59. Miller, G.A. & Friedman, E.A. The reconstruction of mutilated English texts. Information and Control, 1957, 1, 38-55.

60. Miller, G.R. & Coleman, E.B. A set of thirty-six prose passages calibrated for complexity. Journal of Verbal Learning and Verbal Behavior, 1967, 6, 851-854.

61. National Assessment of Educational Progress. Reading objectives. Ann Arbor, Michigan: National Assessment of Educational Progress, 1970.

62. National Assessment of Educational Progress. Report 02-GIY, reading and literature: general information yearbook. Denver: Education Commission of the States, 1972.

63. Powers, R.D., Sumner, W.A., & Kearl, B.E. A recalculation of four adult readability formulas. Journal of Educational Psychology, 1958, 49, 99-105.

64. Pierce, J.R. & Karlin, J.E. Reading rates and the information rate of a human channel. Bell System Technical Journal, 1957, 36, 497-516.

65. Postman, L. & Rosenweig, M.R. Perceptual recognition of words. Journal of Speech Disorders, 1957, 22, 245-253.

66. Rinsland, H.D. A basic vocabulary of elementary school children. New York: Macmillan, 1945.

67. Ruddell, R.B. The effect of the similarity of oral and written patterns of language structure on reading comprehension. Elementary English, 1965, 42, 403-410.

68. Salzinger, K., Portnoy, S., & Feldman, R. The effect of order of approximation to the statistical structure of English on the emission of verbal responses. Journal of Experimental Psychology, 1962, 64, 52-57

69. State Education Department. Educational statistics: New York State. Prepared especially for members of the Legislature. Albany: The State Education Department, January, 1973.

70. Shannon, C.E. A mathematical theory of communication. Bell System Technical Journal, 1948, 27, 379-423, 623-656.

71. Shannon, C.E. Prediction and entropy of printed English. Bell System Technical Journal, 1951, 30, 50-64.

72. Shepard, R.N. Production of constrained associates and the informational uncertainty of the constraint. American Journal of Psychology, 1963, 76, 218-228.

73. Spache, G. A new readability formula for primary grade reading materials. Elementary School Journal, 1953, 53, 410-413.

74. Stauffer, R.G. A vocabulary study comparing reading, arithmetic, health, and science texts. The Reading Teacher, 1966, 20, 141-147.

75. Stolurow, L.M. & Newman, J.R. A factorial analysis of objective features of printed language presumably related to reading difficulty. Journal of Educational Research, 1959, 52, 243-251.

76. Stone, C.R. Measuring difficulty of primary reading material: a constructive criticism of Spache's measure. Elementary School Journal, 1956, 57, 36-41.

77. Szalay, T.G. Validation of the Coleman readability formulas. Psychological Reports, 1965, 17, 965-966.

78. Taylor, W.L. Application of "cloze" and entropy measures to the study of contextual constraint in samples of continuous prose. Unpublished doctoral dissertation, University of Illinois, 1954.

79. Taylor, W.L. "Cloze procedure": a new tool for measuring readability. Journalism Quarterly, 1953, 30, 415-433.

80. Taylor, W.L. "Cloze readability scores as indices of individual differences in comprehension and aptitude." Journal of Applied Psychology, 1957, 41, 19-26.

81. Thorndike, E.L. Reading and reasoning: a study of mistakes in paragraph reading. Journal of Educational Psychology, 1917, 8, 323-332.

82. Thorndike, E.L. & Lorge, I. The teachers word book of 30,000 words. New York: Teachers College, 1944.

83. Thurstone, L.L. Note on a reanalysis of Davis' reading tests. Psychometrika, 1946, 11, 185-188.

84. Vogel, M. & Washburne, C. An objective method of determining grade placement of children's reading material. Elementary School Journal, 1928, 28, 373-381.

APPENDIX

AN INTRODUCTION TO TYPE-TOKEN MATHEMATICS.

THE LOGNORMAL DISTRIBUTION

In this appendix we consider the problem of describing a large vocabulary, such as that of the English language, in mathematical terms. We also address the problem of estimating the true statistical parameters of such a vocabulary from the properties of actual samples.

Notations and Terminology

Let us regard the vocabulary, or lexicon, as a set $\mathfrak{T}$ containing a large number Φ of types (different words). This set $\mathfrak{T}$ is then the total theoretical reservoir from which users of the language must draw in speaking or writing.

Each type T in the set $\mathfrak{T}$ is assumed to have a certain theoretical probability of occurrence, denoted by $\pi(T)$, or simply by π. Thus, in a very large sample, say of N tokens, we should expect very nearly $N\pi$ of these tokens to be instances of the particular type T.

Now let us subdivide the entire lexicon $\mathfrak{T}$ into disjoint classes, on the basis of probability:

$$\mathfrak{T} = \mathfrak{T}_1 \cup \mathfrak{T}_2 \cup \cdots \cup \mathfrak{T}_M \qquad (1)$$

Thus, $\mathfrak{T}_1$ consists of all types having probability π_1, while $\mathfrak{T}_2$ consists of all types having probability π_2, and so on. We may as well suppose that the numbering has been done in order of increasing π, so that π_1 is the smallest, and

π_M the largest.

For each integer i (i=1,2,...,M), let ϕ_i denote the number of types in the class $\mathcal{T}_i$. Then the fraction $\lambda_i = \phi_i/\Phi$ is the proportion of types in that class. We may then sum these λ_i, to obtain the cumulative proportion of types having probabilities $\leq \pi_i$:

$$\Lambda_i = \sum_{K=1}^{i} \lambda_K \qquad (2)$$

Obviously we have $\Lambda_M = 1$. We may also set $\Lambda_o = 0$, so that $\Delta\Lambda_i = \Lambda_i - \Lambda_{i-1} = \lambda_i$ for all i (i=1,2,...,M).

The above expression Λ_i is actually a distribution function, in the sense of general probability theory. To be specific, consider the experiment of choosing a type T at random from the lexicon and observing its probability $\pi(T)$. The value observed is then a random variable, and Λ_i is its distribution function; that is, Λ_i is the probability that the observed probability will not exceed π_i (The double occurrence of the word "probability" in the last sentence is a source of some confusion until one gets accustomed to it.) We shall frequently refer to Λ_i as the type distribution.

There is also a token distribution, corresponding to a different random variable, obtained from a different experiment. This time we select a token at random from the entire written language, and observe the probability of the type represented by the selected token. If we let λ_i^* denote the proportion of tokens accounted for by the types in class $\mathcal{T}_i$, then the token distribution is:

$$\Lambda_i^* = \sum_{K=1}^{i} \lambda_K^* \qquad (3)$$

In this appendix, starred symbols will always refer to the token distribution, while the corresponding unstarred symbols will refer to the type distribution.

The two distributions describe quite different phenomena, as can be seen by considering approximations to the two experiments sketched above. For the first case (the type distribution), we might open the Oxford English Dictionary at random, and select one of the entry-words. For the second case (the token distribution), we might choose a book at random from the Library of Congress, and select any word from it. It should be obvious that the results will differ: the first experiment will most often produce a word from the middle of the probability range, while the second will most often produce a word from the high end.

Some Basic Relationships

While the type and token distributions describe different phenomena, as just explained, they are mathematically related to each other. In fact, either of them can be derived from the other. To see this, consider the class $\mathcal{C}_i$. There are ϕ_i types in this class, and each of these types has probability π_i, which means that <u>each</u> of these ϕ_i types should account for the proportion π_i of the tokens in a large sample. Hence the proportion of tokens accounted for by <u>all</u> the types in $\mathcal{C}_i$ is:

$$\lambda_i^* = \pi_i \ \phi_i \qquad (4)$$

and so we have:

$$\Lambda_i^* = \sum_{K=1}^{i} \pi_K \phi_K$$

$$= \Phi \sum_{K=1}^{i} \pi_K \lambda_K$$

$$= \Phi \sum_{K=1}^{i} \pi_K \Delta\Lambda_K \qquad (5)$$

Thus, a knowledge of the type distribution Λ_i enables one to calculate the token distribution Λ_i^*.

Another relationship of considerable importance is a direct consequence of (5), obtained by setting i=M. We observe that $\Lambda_M^* = 1$, and hence we must have:

$$\Phi = \frac{1}{\sum_{K=1}^{M} \pi_K \Delta\Lambda_K} \qquad (6)$$

This means that the total theoretical vocabulary size can be calculated from knowledge of the type distribution.

We may now combine Eqs. (5) and (6), to obtain an alternate version of (5) which reveals the token distribution to be precisely the so-called "first-order moment-distribution" of the type distribution:

$$\Lambda_i^* = \frac{\sum_{K=1}^{i} \pi_K \cdot \Delta\Lambda_K}{\sum_{K=1}^{M} \pi_K \cdot \Delta\Lambda_K} \qquad (7)$$

The Incidence Numbers

In a sample of N tokens, we may expect a certain number F_1 of types to occur exactly once each, a certain number F_2 of types to occur exactly twice each, a certain number F_3 of types to occur exactly three times each, and so on. We may also expect a certain number F_0 of types not to occur at all in the sample. We shall refer to these numbers F_0, F_1, F_2, ... as the incidence numbers. We now consider the problem of calculating the incidence numbers, assuming we have knowledge of the true theoretical type distribution Λ_i.

Of course, if the sample size N were very large indeed (say, in the trillions or quadrillions), then the problem would be quite simple. Each type would then occur almost exactly as often as its true probability dictates, and so the incidence numbers would coincide with the numbers ϕ_i. In actual practice, however, the sample size N will not be nearly large enough for this simple approach; we must therefore proceed differently.

Consider a particular type belonging to the class $\mathcal{C}_i$. Since this type has true probability π_i, the probability that this type will occur exactly j times in a sample of size N is given by the well-known "binomial" formula:

$$\binom{N}{j} \pi_i^j (1-\pi_i)^{N-j}$$

Now, there are ϕ_i types in class $\mathcal{C}_i$, and so the number of types from class $\mathcal{C}_i$ which may be expected to occur exactly j times in a sample of size N is the product of the preceding expression by the number ϕ_i:

$$\begin{aligned} f_{ij} &= \binom{N}{j} \pi_i^j (1-\pi_i)^{N-j} \quad \phi_i \\ &= \Phi \binom{N}{j} \pi_i^j (1-\pi_i)^{N-j} \quad \Delta\Lambda_i \end{aligned} \tag{8}$$

Finally, the total number of types from <u>all</u> classes which may be expected to occur exactly j times in a sample of size N is obtained by summing (8) over all values of the index i. Thus we may calculate the incidence numbers:

$$\begin{aligned} F_j &= \sum_{i=1}^{M} f_{ij} \\ &= \Phi \binom{N}{j} \sum_{i=1}^{M} \pi_i^j (1-\pi_i)^{N-j} \quad \Delta\Lambda_i \end{aligned} \tag{9}$$

Checking for Internal Consistency

Formula (9) may be checked in two ways. In the first place, the sum of all the incidence numbers F_j should agree with the total number of types:

$$\sum_{j=0}^{N} F_j = \Phi \sum_{i=1}^{M} \sum_{j=0}^{N} \binom{N}{j} \pi_i^{\,j} (1-\pi_i)^{N-j} \Delta\Lambda_i$$

$$= \Phi \sum_{i=1}^{M} \Delta\Lambda_i$$

$$= \Phi \, \Lambda_M$$

$$= \Phi$$

In the second place, the sum of all the products jF_j should agree with the total number of tokens:

$$\sum_{j=0}^{N} jF_j = \Phi \sum_{i=1}^{M} \sum_{j=0}^{N} \binom{N}{j} j\pi_i^j (1-\pi_i)^{N-j} \Delta\Lambda_i$$

$$= \Phi N \sum_{i=1}^{M} \pi_i \Delta\Lambda_i$$

$$= N \sum_{i=1}^{M} \pi_i \phi_i$$

$$= N\Lambda_M^*$$

$$= N$$

In both of the above calculations, we made use of certain well-known tabulated results on sums involving the binomial coefficients.*

The Lognormal Distribution

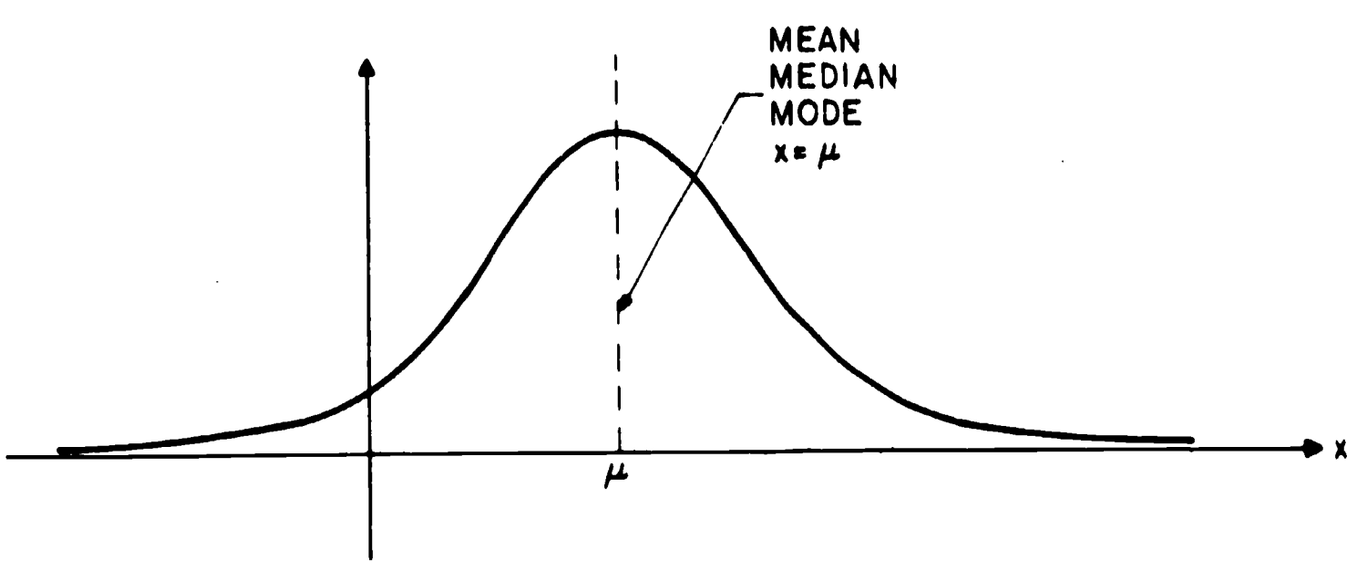
There is considerable evidence to show that the type distribution obtained from any natural language is very closely approximated by a continuous distribution known as the lognormal distribution. Accordingly, we now interrupt our discussion of vocabulary statistics to examine some of the general features of the lognormal distribution.

*See, for example, Handbook of Mathematical Functions, Abramowitz, M. and Stegun, I.A. (eds.), Applied Mathematics Series #55, National Bureau of Standards, 1964, pp. 822-823.

We begin by considering the normal distribution, which is unquestionably the best-known of all the continuous distributions. A random variable X is said to be normally distributed if, for every real number x, the probability that $X < x$ is given by the formula:

$$\text{prob}\left\{X < x\right\} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[\frac{-(t-\mu)^2}{2\sigma^2}\right] dt \qquad (10)$$

The parameters μ and σ which enter into this formula are precisely the mean and the standard deviation of the distribution, respectively. Indeed, because of the symmetry and unimodality of the integrand, the parameter μ is simultaneously the mean, median, and mode of the distribution. The following graph of the integrand exhibits the familiar bell-shape of the "normal" curve:



It is a trivial matter, of course, to shift the origin and change the scale, so that μ becomes 0 and σ becomes 1. When this is done, we say that the random variable X is <u>standardized</u>. The following notations are the usual ones found in the literature:

$$\text{Integrand:} \qquad \mathfrak{z}(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

$$\text{Integral:} \qquad \mathfrak{P}(x) = \int_{-\infty}^{x} \mathfrak{z}(t)\,dt$$

$$\text{Complementary Integral:} \qquad \mathfrak{Q}(x) = \int_{x}^{\infty} \mathfrak{z}(t)\,dt$$

$$= 1 - \mathfrak{P}(x) \quad .$$

We shall have some occasion to refer to the functions $\mathfrak{z}$, $\mathfrak{P}$, and $\mathfrak{Q}$ in what follows.

Now let Y be another random variable, this time limited to positive values only. We say that Y is <u>lognormally</u> distributed if its logarithm $X = \ln Y$ is normally distributed. For a given positive real number y, then, the probability that $Y < y$ is the same as the probability that $X < \ln y$, and is therefore given by the formula (10) with the substitution of $\ln y$ for x as the upper limit of integration:

$$\text{prob}\left\{Y < y\right\} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\ln y} \exp\left[\frac{-(t-\mu)^2}{2\sigma^2}\right] dt$$

If we change the variable of integration by the equation $t = \ln s$, we obtain the following formula for the lognormal

distribution function:

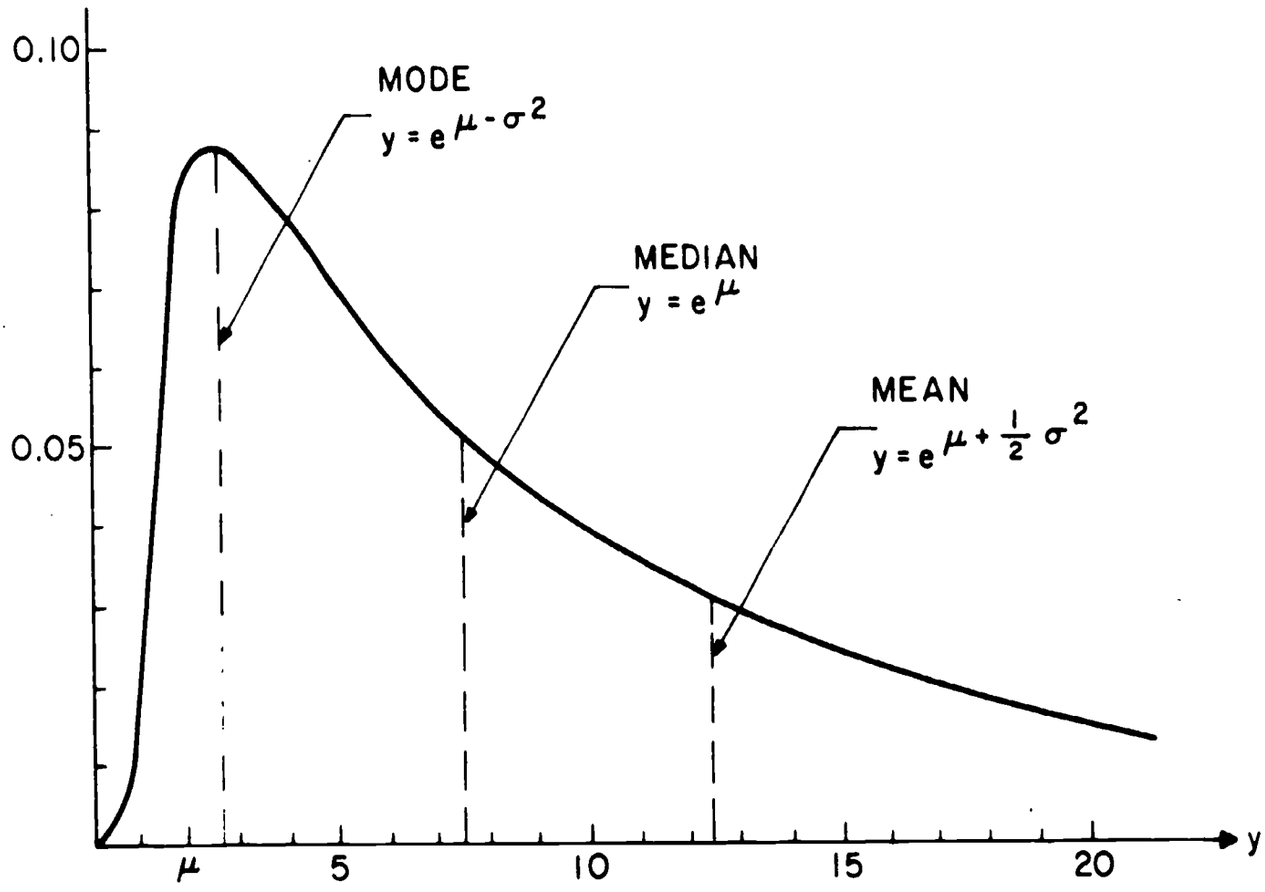
$$\text{prob}\left\{Y < y\right\} = \frac{1}{\sigma\sqrt{2\pi}} \int_0^y \frac{1}{s} \exp\left[\frac{-(\ln s - \mu)^2}{2\sigma^2}\right] ds \quad (11)$$

Here we see that the lognormal distribution also involves two parameters μ and σ, just as the parent normal distribution does. It is important to note, however, that μ and σ no longer have the same interpretation as mean and standard deviation; a bit of calculation shows that for the lognormal distribution (11) these quantities are given by:

$$\text{mean} = e^{\mu + \frac{1}{2}\sigma^2}$$

$$\text{standard deviation} = e^{\mu + \frac{1}{2}\sigma^2} \sqrt{e^{\sigma^2} + 1}$$

and that the median and the mode no longer coincide with the mean. The following graph of the integrand of (11) is drawn for $\mu = 2$, $\sigma = 1$, for the sake of illustration; note the different scales on the horizontal and vertical axes.



Because a lognormally distributed random variable assumes positive values only, it is possible to define moments of arbitrary order (including fractional order) for such a variable. To be explicit, the α-th order moment of the distribution (11) is obtained by inserting the factor s^{α} in the integrand and then integrating over the full range:

$$M_{\alpha} = \frac{1}{\sigma\sqrt{2\pi}} \int_{0}^{\infty} s^{\alpha-1} \exp\left[\frac{-(\ln s - \mu)^2}{2\sigma^2}\right] ds \qquad (12)$$

A bit of simple calculus leads to an exact evaluation of this expression: $M_{\alpha} = \exp\left[\alpha\mu + \frac{\alpha^2\sigma^2}{2}\right]$.

Again because we are dealing with positive values only, it makes sense to define moment-distributions of arbitrary order. These are defined just like the moments above, except that the upper limit of integration becomes variable again; the factor $1/M_\alpha$ is also introduced to normalize the resulting integrals in the conventional manner.

Thus the α-th order moment-distribution corresponding to (11) would be as follows:

$$D_\alpha(y) = \frac{1}{M_\alpha \sigma\sqrt{2\pi}} \int_0^y s^{\alpha-1} \exp\left[\frac{-(\ln s-\mu)^2}{2\sigma^2}\right] ds \qquad (13)$$

In case $\alpha = 0$, the above expression coincides with (11), because $M_0 = 1$. Thus the zeroth order moment-distribution is identically the original distribution.

Now we come to the remarkable self-replicating property which is peculiar to lognormal distributions. Upon performing a few elementary manipulations, we find that (13) can be re-written in the following revealing form:

$$D_\alpha(y) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^y \frac{1}{s} \exp\left[\frac{-(\ln s-\mu-\alpha\sigma^2)^2}{2\sigma^2}\right] ds \qquad (14)$$

Comparison of (14) with (11) shows almost exact agreement; the only difference is that "μ" in (11) is replaced by "$\mu + \alpha\sigma^2$" in (14). This means that every moment-distribution of a lognormal distribution is itself lognormal. Moreover, there is a simple relation between the parameters: the "σ" parameter is unchanged, and the "μ" parameter becomes $\mu + \alpha\sigma^2$.

It will be convenient, if what follows, to have a special notation for the expressions which constantly recur in discussions of lognormal distributions. Accordingly, we introduce the symbols $\mathcal{D}$ and $\mathcal{L}$ for the integrand and the integral of (11), as follows:

$$\text{Integrand:} \quad \mathcal{D}(y,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{y} \exp\left[\frac{-(\ln y-\mu)^2}{2\sigma^2}\right]$$

$$\text{Integral:} \quad \mathcal{L}(y,\mu,\sigma) = \int_0^y \mathcal{D}(s,\mu,\sigma)ds$$

Vocabulary Statistics

Now let us resume our discussion of vocabulary statistics. Our earlier results, embodied in Eqs. (1) through (9), were all exact; but now we shall introduce some approximations which will make the calculations more tractable. The basic assumption is that the type distribution Λ_i is, in fact, very nearly lognormal:

$$\Lambda_i \approx \mathcal{L}(\pi_i,\mu,\sigma) \tag{15}$$

Since we already know that the token distribution Λ_i^* is the first-order moment-distribution of Λ_{i*}, it follows from the self-replication property of $\mathcal{L}$ that Λ_i^* is also very nearly lognormal:

$$\Lambda_i^* \approx \mathcal{L}(\pi_i,\mu + \sigma^2,\sigma) \tag{16}$$

Our problem, of course, is to determine the correct values of the parameters μ and σ from a sampling of words. If

our basic assumption is correct, these two numbers will furnish a very concise and complete description of the entire lexicon from which the sample is drawn.

A few words are in order at this point concerning the nature of the approximation we are using. The true type distribution, after all, is discrete, because the lexicon contains only a finite number of types; yet we are assuming that it is well approximated by the lognormal distribution, which is continuous. It is by no means uncommon to pass from the discrete to the continuous in statistical work, but in this case we may have some problems, especially at the upper end of the frequency (probability) scale.

One difficulty is that the types of highest frequency have widely and irregularly scattered frequencies. This is clearly shown by the seven most common types, whose frequencies (taken from the AHI corpus)* are as follows:

Type	Frequency
the	0.0731
of	0.0285
and	0.0262
a	0.0244
to	0.0236
in	0.0194
is	0.0116

Such departures from continuity indicate that our lognormal model becomes very unsatisfactory for the words which occur

*John B. Carroll et al., Word Frequency Book, American Heritage Publishing Co., New York, 1971.

most frequently.

Another difficulty is that the lognormal curve is unbounded on the right. Since, for our applications, the abscissae represent word-type probabilities, this means that the lognormal model provides for types of arbitrarily high probability (i.e., even greater than 1). Here again the model fails at the upper end.

We have considered resolving these difficulties by changing our model from lognormal to "truncated" lognormal. In other words, we might assume that the type distribution is lognormal up to a certain cut-off point; and beyond that point, it is discrete. There is no doubt that this assumption is more nearly true than that of 100% lognormality. Some preliminary calculations, however, have indicated that there is very little difference in the final results--the values of μ and σ --obtained from the two assumptions. Accordingly, for the remainder of this discussion, we adopt the assumption that type and token distributions are completely lognormal.

The passage from discrete to continuous representations entails a number of notational changes, and it may be well to take note of how these changes affect the appearance of our earlier results. The discrete variable π_i gives way to the continuous variable π, and similarly Λ_i and Λ_i^* become $\Lambda(\pi)$ and $\Lambda^*(\pi)$. The finite sums are replaced by integrals, and the discrete differences $\Delta\Lambda_i$ become differentials $d\Lambda(\pi)$. Note that since we are assuming lognormality we have

$$\Lambda(\pi) = \mathcal{L}(\pi,\mu,\sigma)$$

$$d\Lambda(\pi) = \mathcal{D}(\pi,\mu,\sigma)d\pi$$

With such changes in mind, we may rewrite Eq. (6), which gives a formula for the total vocabulary Φ, as follows:

$$\Phi = \frac{1}{\int_0^1 \pi \, \wp(\pi,\mu,\sigma)\, d\pi} \qquad (17)$$

Note that we have taken the upper limit of integration to be 1; there is no point in making it any higher, since the variable π represents a probability. In the same way our formula (9) for the incidence numbers may be rewritten as:

$$F_j = \Phi\binom{N}{j}\int_0^1 \pi^j (1-\pi)^{N-j} \wp(\pi,\mu,\sigma)\, d\pi \qquad (18)$$

Calculation of μ and σ

Now that we have a grasp of the underlying mathematical theory, we can proceed to the problem of determining the values of μ and σ from the properties of an actual sample. Let the sample consist of N tokens altogether, and let us suppose that all the preliminary processing (such as lemmatization and resolution of ambiguities) has been completed.

We can then determine the sample incidence numbers G_1, G_2, G_3, ... by a straightforward count. Thus, G_1 is the number of types which are found to occur once each in the sample, G_2 is the number of types which are found to occur twice each in the sample, and so on. Note the distinction between the sample incidence numbers G_j and the theoretical incidence numbers F_j previously mentioned:

F_j is the number of types which may be expected to occur exactly j times each in a sample of size N, assuming knowledge of the true type distribution.

G_j is the number of types which actually did occur exactly j times each in the sample.

Note also that we have no direct way of observing the value of G_o, the number of types which failed to occur at all in the sample.

Next we begin an iterative calculation; we simply guess at the values of μ and σ to start the process. Naturally enough, the closer our initial guesses are to the true values, the more rapidly the iteration will converge. We could, for example, derive our initial values from the values already obtained in earlier vocabulary studies (e.g. the AHI corpus), after estimating the effect of lemmatization.*

From the assumed values of μ and σ we can calculate the total theoretical vocabulary Φ by Eq. (17). This gives us a way to estimate the missing sample incidence number G_o, since:

$$G_o = \Phi - \sum_{j \geq 1} G_j \qquad (19)$$

We can also calculate the theoretical incidence numbers F_j (j=0, 1, 2, 3, ...) by Eq. (18). This calculation itself presents some interesting problems. These problems are explained in the final section of this appendix.

Next, we compare the theoretical incidence numbers F_j with the actual sample incidence numbers G_j. If the two families of numbers agree substantially, then we are satisfied that the assumed values of μ and σ were correct, and we are finished. If not, then we must choose new values for μ and σ, and repeat the process.

---

*Lemmatization is discussed in Chapter II.

Two problems arise here. First, we must decide exactly what constitutes "substantial agreement" between the two families of numbers. Second, when this agreement is not achieved, we must know how to modify the values of μ and σ so that the next "pass" will be more nearly successful.

There are many ways to dispose of these two problems. We might very well adopt the procedure used by John B. Carroll in his statistical analysis of the AHI corpus. In broad outline, this procedure was as follows:

1. Define two functions $M(X_0, X_1, X_2, \ldots)$ and $S(X_0, X_1, X_2, \ldots)$. The exact choice of definition made by Carroll is somewhat elaborate, and need not concern us here. Suffice it to say that M and S are the values of the sample parameters corresponding to μ and σ, for a given set of incidence numbers $X_0, X_1, X_2, \ldots$.

2. Evaluate both functions, first with the theoretical incidence numbers $F_0, F_1, F_2, \ldots$ as arguments, and second with the sample incidence numbers $G_0, G_1, G_2, \ldots$ as arguments.

3. If the evaluations agree, say to three decimal places, we are finished. If not, choose new parameters μ' and σ' by the "correction" equations:

$$\mu' = \frac{M(G_0, G_1, G_2, \ldots)}{M(F_0, F_1, F_2, \ldots)} \mu$$

$$\sigma' = \frac{S(G_0, G_1, G_2, \ldots)}{S(F_0, F_1, F_2, \ldots)} \sigma$$

(Note that the particular form of these correction equations is a result of the way the functions M and S were defined).

4. Using the new values μ' and σ' (instead of μ and σ), recalculate the theoretical incidence numbers F_0, F_1, F_2, ... (and also G_0), and return to step 2.

This type of procedure, however, is not the only one available. The literature of numerical analysis abounds with methods of achieving convergence in an iterative calculation. The choice of method should be left open until the actual data are available.

Calculation of Incidence Numbers

The actual calculation of the theoretical incidence numbers F_j by Eq. (18) presents some formidable problems in numerical analysis:

$$F_j = \Phi \binom{N}{j} \int_0^1 \pi^j (1-\pi)^{N-j} \mathcal{J}(\pi, \mu, \sigma) d\pi$$

$$= \frac{\Phi}{\sigma\sqrt{2\pi}} \binom{N}{j} \int_0^1 \pi^{j-1} (1-\pi)^{N-j} \exp\left[\frac{-(\ln \pi -)^2}{2\sigma^2}\right] d\pi \qquad (20)$$

The chief problem is the factor $(1-\pi)^{N-j}$, which varies from 1 to 0 in the interval of integration. Because of the immense size of N, it does so with great speed: this means that almost all of the integral's value is contributed by the values of π at the extreme left end of the interval. In this region the integrand varies so rapidly as to defy ordinary methods of numerical quadrature.

One solution, which has been tested and which seems to work very well, consists of approximating the integral (20) by a sum of integrals over subintervals. These subintervals are so chosen that the troublesome factor $(1-\pi)^{N-j}$ is es-

sentially constant in each of them. Specifically, we proceed as follows:

1. Choose a small positive number ϵ, say $\epsilon = 1/M$ where M is an integer.

2. Let $H_0 = 1$, $H_1 = 1-\epsilon$, $H_2 = 1-2\epsilon$, and so on. That is, $H_i = 1-i\epsilon$. Then we have

$$1 = H_0 > H_1 > H_2 > \ldots > H_M = 0$$

3. Define the numbers π_i by setting $(1-\pi_i)^{N-j} = H_i$. That is, $\pi_i = 1 - (H_i)^{1/N-j}$. Since the H_i form a decreasing sequence, the π_i must form an increasing one:

$$0 = \pi_0 < \pi_1 < \pi_2 < \ldots < \pi_M = 1$$

4. Now let S_i denote the same integral as (20), but taken only over the i-th subinterval defined by the π's:

$$S_i = \frac{\Phi}{\sigma\sqrt{2\pi}} \binom{N}{j} \int_{\pi_{i-1}}^{\pi_i} \pi^{j-1}(1-\pi)^{N-j} \exp\left[\frac{-(\ln \pi-\mu)^2}{2\sigma^2}\right] d\pi$$

The desired quantity F_j is just the sum of these S_i's:

$$F_j = \sum_{i=1}^{M} S_i$$

5. Let T_i be the same as S_i, but without the factor $(1-\pi)^{N-j}$ in the integrand:

$$T_i = \frac{\Phi}{\sigma\sqrt{2\pi}}\binom{N}{j}\int_{\pi_{i-1}}^{\pi_i} \pi^{j-1} \exp\left[\frac{-(\ln \pi-\mu)^2}{2\sigma^2}\right]$$

It happens to be a straightforward exercise to calculate T_i explicitly in terms of the standard normal distribution function $\mathcal{P}(X)$. The result is:

$$T_j = \Phi\binom{N}{j} \exp\left[j\mu + \frac{j^2\sigma^2}{2}\right] (P_i - P_{i-1})$$

$$\text{where} \quad P_i = \mathcal{P}\left[\frac{\ln \pi_i-\mu-j\sigma^2}{\sigma}\right]$$

6. Obviously we have

$$H_iT_i < S_i < H_{i-1}T_i$$

and so, summing over i, we obtain:

$$\sum_{i=1}^{M} H_iT_i < F_j < \sum_{i=1}^{M} H_{i-1}T_i$$

7. The two sums which flank the preceding inequality are readily calculated. To do this, let us take out the leading factor of T_i (that is, Φ times the binomial coefficient times the exponential), which is independent of i, and consider the following sums:

$$\sum_{i=1}^{M} H_i(P_i - P_{i-1})$$

and

$$\sum_{i=1}^{M} H_{i-1}(P_i - P_{i-1})$$

A bit of manipulation ("summation by parts"), together with knowledge of the regular spacing of the H_i, allows us to write these sums in a particularly simple form:

$$\sum_{i=1}^{M} H_i(P_i - P_{i-1}) = \epsilon \sum_{i=1}^{M-1} P_i$$

$$\sum_{i=1}^{M-1} H_{i-1}(P_i - P_{i-1}) = \epsilon \sum_{i=1}^{M} P_i$$

8. The quantities P_i required here are easily found, because they are simply values of the standard normal distribution function $\mathcal{P}(x)$. This function is well approximated by the first few terms of its Maclaurin series for small values of x, while for larger values we can use a rapidly convergent continued-fraction expansion.

Thus we have a non-tentative numerical procedure for calculating the F_j. To summarize it, we have:

$$F_j = \Phi \binom{N}{j} \exp\left[j\mu + \frac{j^2\sigma^2}{2}\right] \epsilon \sum_{i=1}^{M} P_i \qquad (21)$$

where the relative error is less than $\frac{P_M}{M}$ (which can be made as small as desired by increasing the size of M).

$$\sum_{i=1} P_i$$

There is need for considerable care, of course, in evaluating (21), especially for the larger values of j. This is because the leading factor (the binomial coefficient and the exponential) increases very rapidly with j to astronomical size, while the trailing factor (the sum of the P_i's) decreases with comparable speed. Treating the two factors separately would quickly lead to intolerable "noise" in the calculation. Once this is recognized, however, it is straightforward to keep this situation under control.